Recent Advances of Foundation Language Models-based Continual Learning: A Survey

Yutao Yang[1], Jie Zhou*[1], Xuanwen Ding[1], Tianyu Huai[1], Shunyu Liu[1], Qin Chen[1] and Liang He[1]

[1] *East China Normal University, The School of Computer Science, Shanghai, China*

Reception date of the manuscript: dd/mm/aaaa
Acceptance date of the manuscript: dd/mm/aaaa
Publication date: dd/mm/aaaa

Abstract— Recently, foundation language models (LMs) have marked significant achievements in the domains of natural language processing (NLP) and computer vision (CV). Unlike traditional neural network models, foundation LMs obtain a great ability for transfer learning by acquiring rich commonsense knowledge through pre-training on extensive unsupervised datasets with a vast number of parameters. However, they still can not emulate human-like continuous learning due to catastrophic forgetting. Consequently, various continual learning (CL)-based methodologies have been developed to refine LMs, enabling them to adapt to new tasks without forgetting previous knowledge. However, a systematic taxonomy of existing approaches and a comparison of their performance are still lacking, which is the gap that our survey aims to fill. We delve into a comprehensive review, summarization, and classification of the existing literature on CL-based approaches applied to foundation language models, such as pre-trained language models (PLMs), large language models (LLMs) and vision-language models (VLMs). We divide these studies into offline CL and online CL, which consist of traditional methods, parameter-efficient-based methods, instruction tuning-based methods and continual pre-training methods. Offline CL encompasses domain-incremental learning, task-incremental learning, and class-incremental learning, while online CL is subdivided into hard task boundary and blurry task boundary settings. Additionally, we outline the typical datasets and metrics employed in CL research and provide a detailed analysis of the challenges and future work for LMs-based continual learning.

Keywords—Continual Learning, Foundation Language Models, Pre-trained Language Models, Large Language Models, Vision-Language Models, Survey

I. Introduction

Recent advancements in foundation language models (LMs) have set new benchmarks in both natural language processing (NLP) [1, 2, 3] and computer vision (CV) [4]. Foundation LMs encompass three primary categories: Pre-trained Language Models (PLMs) [2], Large Language Models (LLMs) [1], and Vision-Language Models (VLMs) [5]. PLMs such as BERT [6], RoBERTa [7], and BART [8] focus on text-based tasks and are crucial for understanding and generating language by leveraging tasks like masked language modeling during pre-training. LLMs, including models like GPT-4 [9] and LLaMA [10], extend the capabilities of PLMs by increasing the scale of model architecture and training data, thus enhancing their generality and adaptability across a broader range of tasks. VLMs, represented by VisualBERT [11], CLIP [12], LLaVA [13] and DALL-E [14], integrate text and image modalities to enable complicated interactions between visual and textual information. The underlying paradigm of these models involves pre-training on extensive, often unlabeled datasets to capture rich semantic information, which is subsequently fine-tuned for specific tasks or domains. This methodology not only boosts performance across various applications but also significantly enhances the models' flexibility and task adaptability.

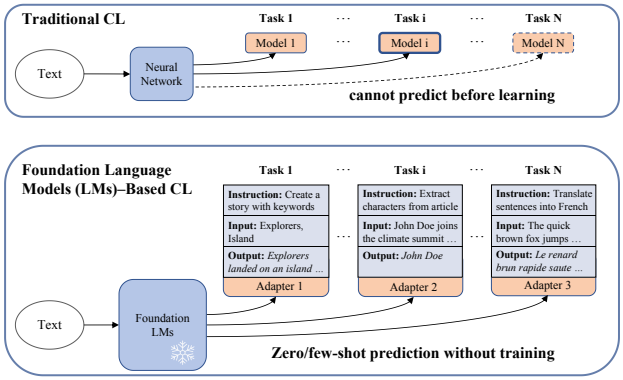

Fig. 1: Comparison between traditional CL and Foundation language models (LMs)-Based CL.

However, these foundation models often demonstrate limitations in dynamic environments with a sequence of tasks, primarily due to their fixed parameters once training is completed. These models generally lack the capability to inte-

Contact data: Jie Zhou, jzhou@cs.ecnu.edu.cn

grate new data or concepts without undergoing a retraining process. A significant challenge associated with training on a sequence of tasks is "catastrophic forgetting" [15], a phenomenon where a model loses previously acquired knowledge upon learning new information. This is in stark contrast to human learning processes, which are inherently continuous and adaptive. Despite the successes of multi-task learning (MTL) and transfer learning (TL) in certain applications, they have limitations in real-world scenarios. MTL necessitates having all tasks and their data available upfront, which poses a challenge when launching a new service as the model must be retrained with all the data. Furthermore, TL is typically done with only two tasks, i.e., the source and the target, rendering it impractical for real-world online platforms with multiple target tasks. To address these challenges, it is crucial for models to process and learn the continuously expanding and diversifying datasets. This requires mechanisms that allow models to adapt to new linguistic phenomena and trends without compromising the accuracy and sensitivity towards historical data.

Consequently, continual learning (CL) [16, 17], also referred to as lifelong learning [18] or incremental learning [19], is a crucial area in artificial intelligence that seeks to develop systems capable of continuously updating themselves and acquiring new knowledge, without forgetting previously learned information, similar to human learning [20]. This paradigm is particularly relevant in the context of foundation language models (LMs), which are challenged by specific issues such as catastrophic forgetting (CF) and cross-task knowledge transfer (KT). Catastrophic forgetting represents a significant challenge, where a model tends to lose previously acquired knowledge upon learning new information. To address this, language models must maintain a robust grasp of past language data while adapting to new linguistic trends. Furthermore, cross-task knowledge transfer is essential for enhancing the continual learning process. Effective KT not only accelerates the learning curve for new tasks (forward transfer) but also enhances the model's performance on prior tasks via the feedback of new knowledge (backward transfer).

Recent advancements in continual learning methodologies have substantially enhanced the adaptability and knowledge retention capabilities of foundational language models (LMs). These developments are crucial for addressing complex challenges previously observed in CL. Researchers have formulated innovative strategies to mitigate these challenges, thereby enabling LMs to maintain high performance across a variety of tasks while continually integrating new knowledge [95, 32, 96]. Notable successes have been documented in diverse downstream tasks, such as aspect-based sentiment analysis, where continual learning enables dynamic adaptation to evolving aspects and sentiments [27]. Similarly, in dialogue generation, the novel technologies assist models in refining and expanding their conversational capabilities through ongoing interactions [97]. In text classification, continual learning facilitates the incorporation of new categories and adjustments to shifts in text distributions without the need for complete retraining [47]. Moreover, in the realm of visual question answering, continual learning is essential for updating the models' capabilities to process and respond to new types of visual content and queries [46, 72]. The aforementioned works underscore the potential of continual learning to significantly boost the performance of foundation LMs.

In the domain of continual learning, there has been a significant paradigm shift from traditional methodologies to those that integrate foundation LMs (See Figure 1). First, foundation LMs demonstrate enhanced generalization and transfer learning abilities across diverse tasks owing to their broad pre-training on large-scale datasets. The model has specialized transfer capability to quickly adapt to downstream tasks with only a few samples. Consequently, it is crucial to mitigate the degradation of both the zero-shot transfer and history task abilities in LMs while facilitating the acquisition of new skills. Second, due to the substantial number of parameters in foundation LMs, it is crucial to employ parameter-efficient techniques [98], such as prompt tuning [99] and adapters [100], to update parameters without comprehensive retraining. Third, the foundation LMs possess the capability to follow instructions through instructional learning [101, 102], enabling more dynamic and context-aware interactions.

This review systematically categorizes these strategies and technologies into two core areas: offline continual learning and online continual learning (Figure 2). We first give detailed definitions and scenarios to format the setting of offline and online CL, where offline CL includes domain-incremental, task-incremental and class-incremental CL, and online CL includes hard task boundary and blurry task boundary. These learning strategies are further subdivided into methods based on Pre-trained Language Models (PLMs), Large Language Models (LLMs), and Vision-Language Models (VLMs). Then, we summarize the related papers about traditional methods, continual pre-training methods, parameter-efficient tuning methods and instruction-based methods. Finally, we static the main datasets from various perspectives and review the key metrics to evaluate the forgetting and transferring of the models.

The main contributions of this survey paper can be summarized as follows.

- We thoroughly review the existing literature on foundation LMs-based CL approaches, which integrate foundation LMs with CL to learn new knowledge without retraining the models. It is quite different from traditional CL since foundation LMs have great abilities of transfer learning, zero-shot and instruction following with huge parameters.
- We give the definitions of different settings and categorize these studies into various classes to better understand the development of this domain. In addition to the traditional methods like replay, regularization and parameter-isolation-based algorithms, we also summarize the works about continual pre-training methods, parameter-efficient tuning methods and instruction tuning-based methods.
- We provide the characters of existing datasets for CL and present the main metrics to evaluate the performance of preventing forgetting and knowledge transfer.
- We discuss the most challenging problems of foundation LMs-based CL and point out promising future research directions in this field.

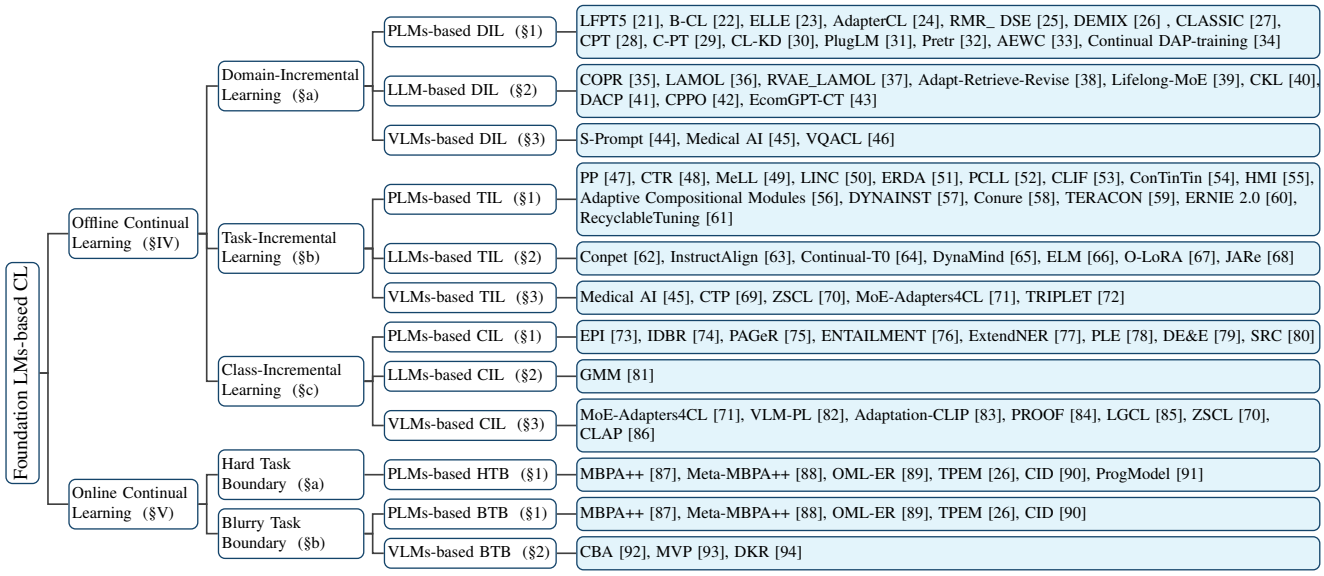

Fig. 2: Taxonomy of foundation language models for continual learning.

The paper is organized as follows. In Section II, we review the mainly related surveys about continual learning. Then, we introduce the base settings and learning modes of continual learning in Section III, including the definitions and scenarios of CL. Furthermore, we present the related studies about offline continual learning, which can be divided into domain-incremental learning, task-incremental learning and class-incremental learning in Section IV. In Section V, we focus on online continual learning, including hard task boundary and blurry task boundary settings. The typical datasets and metrics are provided in Section VI and VII. Finally, we analyze the challenge and further work in Section VIII and give the conclusion in Section IX.

II. Related Surveys

a. Continual Learning

Early examinations of Continual Learning (CL) have provided broad coverage, as observed in surveys such as Parisi et al. [18]. Recently, the Wang et al. [16] conducts a comprehensive survey that categorizes five key strategies in CL: regularization-based, replay-based, optimization-based, representation-based, and architecture-based approaches. This survey reflects an effort to organize and understand the diverse methodologies employed in the field. Notably, there is a growing focus on class-incremental setting [19, 103, 104] and replay-based approaches [105], reflecting the increasing granularity of research interests within the CL domain.

b. Continual Learning for Computer Vision

In the realm of computer vision, De et al. [20] address the pressing challenge of continual learning, specifically in the task-incremental setting, where tasks arrive sequentially with clear boundaries. They introduce a novel stability-plasticity trade-off framework tailored for continual learners and undertake an comprehensive experimental analysis, comparing the efficacy of 11 methodologies across three benchmarks. Qu et al. [106] present a comprehensive examination of continual learning, highlighting its vital role in the accumulation of knowledge from sequential data streams. This research investigates a range of approaches, encompassing regularization, knowledge distillation, memory-based methods, and more. These approaches are systematically categorized by their characteristics and applications in computer vision.

Moreover, Mai et al. [107] focus on the realm of online continual learning within image classification, addressing catastrophic forgetting. This study evaluates the efficacy of state-of-the-art methodologies across diverse memory and data configurations. Masana et al. [103] present a comprehensive performance evaluation of class-incremental methods applied to image classification. The experiments span a range of scenarios, incorporating large-scale datasets and varied network architectures. Belouadah et al. [104] pay more attention to class-incremental learning algorithms for visual tasks. This study defines the essential properties of incremental learning algorithms, offers a unified formalization of the class-incremental learning problem, and provides a evaluation framework for detailed analysis.

c. Continual Learning for NLP

Biesialska et al. [108] address the challenge of continual learning within Natural Language Processing (NLP), wherein conventional architectures struggle to accommodate new tasks without compromising previously acquired knowledge. The research presents an extensive review of methods, including rehearsal, regularization, and architectural approaches, all designed to mitigate the aforementioned challenge.

In a similar vein, Ke et al. [109] offer a focused survey on continual learning within the NLP domain, providing a comprehensive examination of various continual learning settings, methodologies, and challenges. This work presents an in-depth analysis of state-of-the-art approaches and extends original CL settings to be more general and up-to-date. Additionally, it emphasizes the significance of knowledge transfer within NLP and the challenges posed by inter-task class separation.

d. Continual Learning for Other Domains

Recent surveys, surveys like [110, 111, 112] investigate the advancements in incremental learning for neural recommendation systems and continual learning (CL) in robotics, respectively. Zhang et al. [110] make a notable contribution to narrow the gap between academic research and industrial applications through the introduced Incremental Update Recommendation Systems (IURS). They highlight the imperative for real-time updates with streaming data and focus on the distinctive challenges posed by IURS in contrast to traditional Batch Update Recommendation Systems (BURS) and offer a thorough examination of existing literature and evaluation methodologies in this domain. Shaheen et al. [111] provide a comprehensive overview of contemporary approaches for CL within real-world contexts. Their analysis focuses on learning algorithms that efficiently handle large sequential datasets within computational and memory constraints. The survey also explores challenges in applying CL to autonomous systems, comparing methods across metrics such as computational efficiency, memory utilization, and network complexity. In the field of robotics, it is essential for agents to adapt and interact with their environment using a continuous stream of observations. Thus, Lesort et al. [112] explore CL within this domain, defining CL as a paradigm where both data distribution and learning objectives evolve dynamically. They emphasize the challenges in evaluating CL algorithms in robotic applications and introduce a novel framework alongside metrics tailored to effectively present and assess CL methodologies.

This paper centers on the crucial advancements in CL as applied to foundational language models, which have obtained significant success in the fields of NLP and multimodal. We categorize existing works into offline CL and online CL based on pre-trained language models(PLMs), large language models(LLMs), and vision-language models(VLMs).

III. Settings and Learning Modes of CL

a. Basic Formulation

Continual learning is an advanced method in machine learning. Within this framework, the model is sequentially trained across a diverse array of tasks denoted as t within the set $T = \{1, 2, ..., N\}$, where each task t is associated with its individual dataset $X_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{|X_t|}$. Here, $x_i^{(t)}$ represents an individual training example, and $y_i^{(t)}$ denotes the corresponding class label for task t, while $|X_t|$ indicates the total number of samples in task t. However, the data distributions between any two tasks t and t' are distinct ($p(X_t) \neq p(X_{t'})$ for all $t \neq t'$). This distinction presents a fundamental challenge in managing the diversity of data distributions across multiple tasks. This setup necessitates the model to learn new knowledge while retaining past information.

Continual learning encompasses two principal paradigms: offline and online continual learning. These paradigms define how data arrives and how the model updates its knowledge over time.

- Offline Continual Learning: This setting involves learning across a series of tasks, with each task fully presented before handling the next task. For each task t, the model trains on the entire dataset D_t through multiple epochs. The model progresses to task $t+1$ only upon achieving the desired proficiency on task t.
- Online Continual Learning: This setting operates within a dynamic framework wherein the model learns knowledge from a stream of data points or mini-batches presented sequentially. Additionally, the model lacks access to the entire dataset for a given task. This setting closely mirrors real-world scenarios characterized by continuous data flow, compelling the model to adapt in real time.

b. Typical Scenarios

1. Offline Continual Learning

Offline continual learning (See Figure 3) comprises three principal scenarios, each distinguished by distinct characteristics: Domain-Incremental Learning, Task-Incremental Learning, and Class-Incremental Learning.

- Domain-Incremental Learning (DIL): In this scenario, the model aims to process diverse data distributions. Specifically, in DIL, while the data distributions $p(X_t)$ in task t and $p(X_{t'})$ in task t' are different, their task types and class labels remain consistent. The task identities (task-IDs) are not required.
- Task-Incremental Learning (TIL): In this scenario, the model is designed to handle a series of tasks, each with unique objectives. The classes within these tasks may or may not be disjoint. The boundaries of each task are clear, and task-IDs are provided during both training and testing phases.
- Class-Incremental Learning (CIL): In this scenario, the model is designed to continually learn new classes information, while retaining knowledge of previously learned classes. For tasks t and t', while they might share the same task type (such as classification), their class sets C_t and $C_{t'}$ are distinct. Moreover, the task-IDs are only available during training.

In summary, Domain-Incremental Learning concentrates on adapting the model to the shifts in input data distributions while maintaining consistency in tasks and classes. Task-Incremental Learning necessitates the model's ability to learn and retain task-specific knowledge over successive tasks. On the other hand, Class-Incremental Learning highlights the gradual integration of new classes into the model's recognition capabilities without compromising knowledge of previously learned classes.

2. Online Continual Learning

In online continual learning (See Figure 4), existing research is categorized into two configurations based on the arrival pattern of tasks: "Hard Task Boundary" and "Blurry Task Boundary":

- Hard Task Boundary: The arrival of tasks follows a strictly structured and sequential process. Data from the preceding task is completely processed before transitioning to the

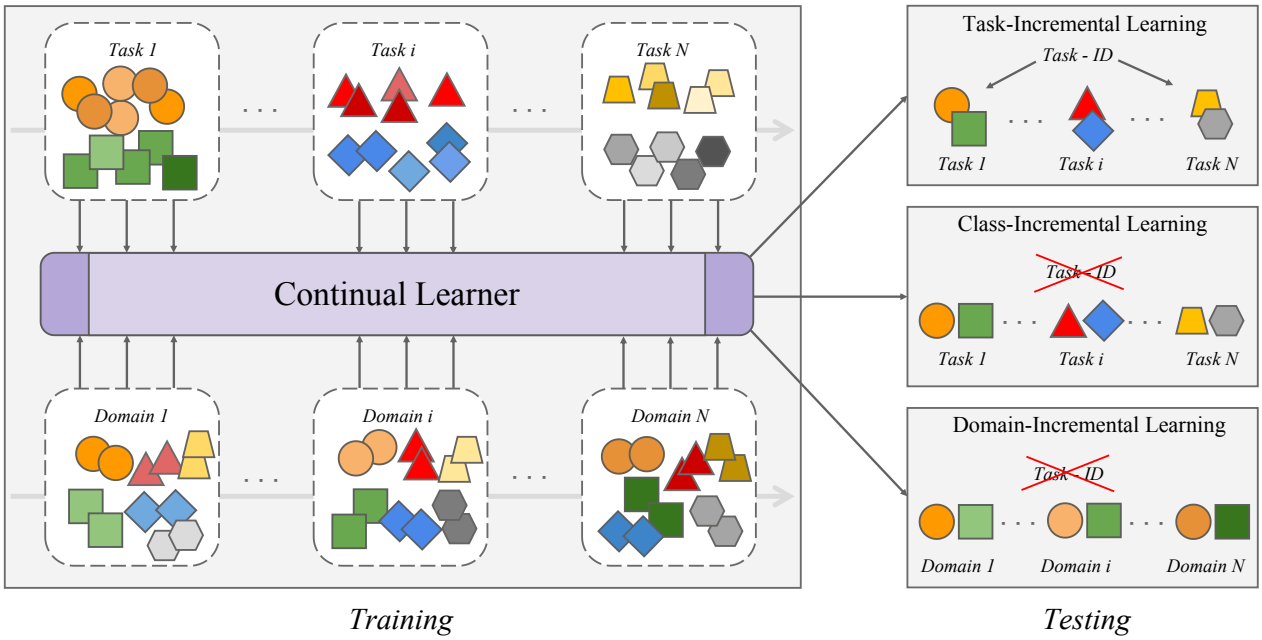

Fig. 3: The setting of different offline continual learning tasks, including task-incremental learning, class-incremental learning and domain-incremental learning. The samples with different classes (domains) are marked with various shapes (colors).

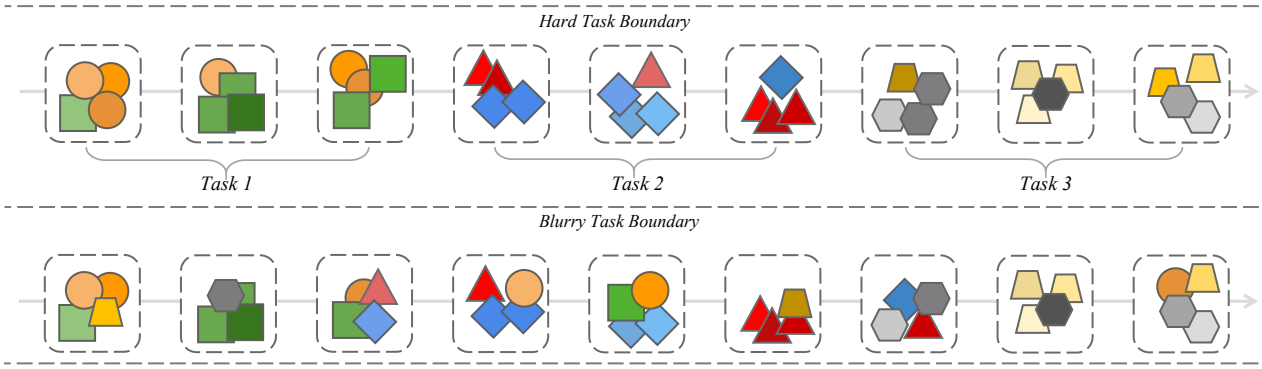

Fig. 4: The setting of different online continual learning tasks, including hard task boundary arriving and blurry task boundary arriving. The samples with different classes (domains) are marked with various shapes (colors).

next task, ensuring no mixing or overlap of data between tasks.

- Blurry Task Boundary: The distinction between tasks is less clear, similar to the real-world scenarios. Data from different tasks are intermixed, making it difficult to pinpoint when one task ends and another begins.

In both setups, the main challenge lies in achieving the balance of learning new data while preserving previously gained knowledge, often termed as catastrophic forgetting. Numerous approaches, such as experience replay [23, 36], elastic weight consolidation (EWC) [15], and progressive neural networks [22, 48], have emerged to address this issue. Each method comes with its unique strengths and weaknesses upon the task arrival configuration.

IV. Offline Continual Learning

Our review of existing literature on offline continual learning identifies three principal categories: domain-incremental learning, task-incremental learning, and class-incremental learning.

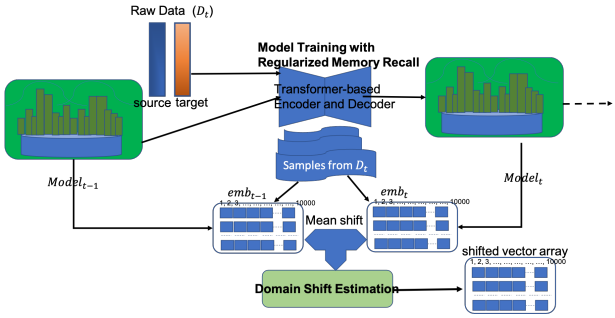

Fig. 5: The framework of RMR_DSE, taken from [25].

a. Domain-Incremental Learning

1. PLMs-based DIL

Traditional Methods. Continual Learning methodologies are frequently employed in the context of Pre-trained Language Models (PLMs), encompassing approaches such as replay-based, regularization-based, and parameter-isolation-based algorithms. Li et al. [25] present a regularization-centric framework for Lifelong Learning (LLL) termed RMR-DSE, specifically tailored for sequential operation

across multiple domains (Figure 5). Unlike conventional strategies requiring incremental memory allocation, RMR-DSE employs a recall optimization mechanism, driven by regularization, to selectively retain important parameters from preceding tasks. Furthermore, it incorporates a domain drift estimation algorithm to address embedding space shifts. Aiming at updating models to learn new languages while maintaining consistent performance across previously learned ones, Castellucci et al. [30] propose a Knowledge Distillation-based Continual Learning (CL-KD) approach. This approach employs a Teacher-Student framework. When the student model is trained on a new language, the teacher model also transfers its knowledge of supported languages to the student model. Lee et al. [33] introduce a domain-agnostic framework. The framework employs a two-phase training process, initially using synthetic data to learn general conversational patterns and subsequently using human-computer dialogs in customer support. Moreover, the Adaptive Elastic Weight Consolidation (AEWC) algorithm is applied to adjust the loss function, ensuring the balance between acquiring new knowledge and retaining previously learned information.

Gururangan et al. [113] develop a parameter-isolation-based method, named DEMIX layers, which consists of a collection of experts. The dynamic addition or removal of these experts can enhance the model's ability to adapt to new domains while maintaining robust performance in previously established ones. PlugLM, devised by Cheng et al. [31], is pre-training model equipped with a differentiable plug-in memory (DPM) designed for domain-adaptive continual training. The core concept behind this approach is to separate the knowledge storage from model parameters using an adaptable key-value memory structure. It enables the explicit retrieval and utilization of knowledge stored within the DPM.

Continual Pre-training Methods. Continual domain-adaptive pre-training (Continual DAP-training) [34] is based on two main ideas: (1) the general knowledge in the LM and the knowledge gained from prior domains are crucial to mitigating CF and enhancing cross-task knowledge transfer (KT). This is achieved through soft-masking units based on their importance, and (2) the model is designed to develop complementary representations of both the current domain and prior domains, thereby facilitating the integration of knowledge. The key novelty of Continual DAP-training is a soft-masking mechanism that directly controls the update to the LM. Cossu et al. [32] formalize and explore the dynamics of continual pre-training (Pretr) scenarios across language and vision domains. In this framework, models undergo continuous pre-training on a sequential stream of data before subsequent fine-tuning for various downstream tasks.

Parameter-Efficient Tuning Methods. Due to the huge parameters of LMs, parameter-efficient tuning methods like adaptors [114, 115] and p-tuning [99] are used for domain-incremental CL [24, 22, 27, 29].

The adapter architecture, as depicted in Figure 6, incorporates a skip-connection to minimize the number of parameters. A notable exemplar of this approach is AdapterCL [24], which employs residual adapters tailored specifically

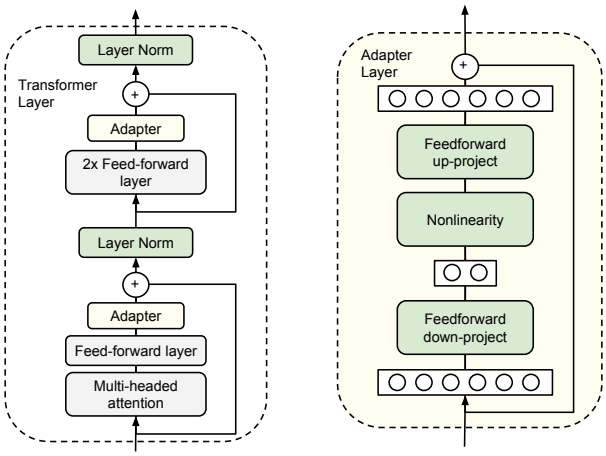

Fig. 6: The structure of Adapter, taken from [114].

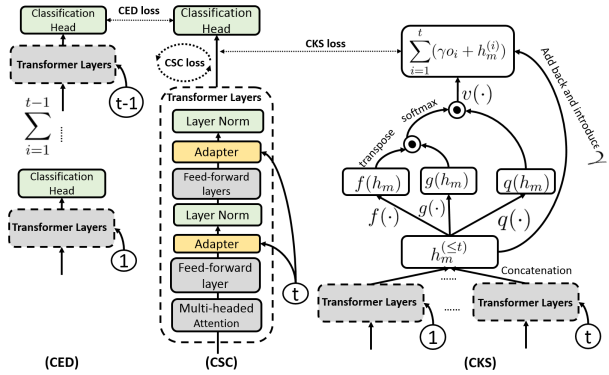

Fig. 7: The framework of CLASSIC, taken from [27].

for task-oriented dialogue systems. This framework, comprising 37 domains, is structured to facilitate continual learning across four important aspects: intent recognition, state tracking, natural language generation, and end-to-end processing. In a related vein, Ke et al. [22] introduce B-CL model to tackle critical challenges in CL for Aspect-Based Sentiment Classification (ABSC). B-CL integrates continual learning adapters within capsule network architectures, enabling B-CL to effectively facilitate knowledge transfer while preserving task-specific information.

Aiming at mitigating catastrophic forgetting during fine-tuning, the CLASSIC model [27] presents an innovative solution by deploying adapters to tap into BERT's capabilities(Figure 7). A novel contrastive continual learning strategy is used to facilitate the transfer of knowledge across tasks and distill insights from previous tasks to subsequent ones. It also effectively eliminates the necessity for task identifiers during testing. Furthermore, Continual PostTraining (CPT) [28] introduces two continual learning plug-in modules, termed CL-plugins, embedded within each transformer layer of RoBERTa. The neurons which is crucial for previous tasks are protected by using the task masks

Prompt tuning [99], or P-tuning, introduces trainable continuous prompts into the sequence of input word embeddings, while the language model remains frozen. To tackle the challenge of CL under limited labeled data, Qin et al. [21] propose a Lifelong Few-shot Language Learning framework (LFPT5). In this framework, prompt tuning, replay and regularization strategies are leveraged. When presented with a new task, the model generates pseudo-labeled samples rep-

resentative of prior domains. The training process then incorporates these pseudo-labeled samples alongside new task-specific data. Additionally, the KL divergence loss is employed to maintain label consistency between the previous and the current model.

Furthermore, Zhu et al. [29] introduced Continual Prompt Tuning (C-PT) as a methodology to address the challenges of continual learning within dialogue systems. C-PT facilitates knowledge transfer between tasks through continual prompt initialization, query fusion, memory replay, and a memory-guided technique.

Instruction Tuning-based Methods. Instruction tuning-based methods involve transforming a given task into natural language instructions. Qin et al. [23] propose ELLE, a novel approach aimed at effectively incorporating continuously expanding streaming data into pre-trained language models (PLMs). It consists of two fundamental components: (1) function-preserved model expansion, which enhances knowledge acquisition efficiency by changing the width and depth of an existing PLM , and (2) pre-trained domain prompts, which significantly enhance the adaptation for downstream tasks by effectively segregating the diverse knowledge acquired during pre-training phases.

2. LLMs-based DIL

In this section, we mainly review the related studies of LLMs-based domain-incremental learning.

Traditional Methods. In many practical scenarios, retraining Language Models (LMs) is challenging due to resource constraints and data privacy concerns. Zhang et al. [42] introduce Continual Proximal Policy Optimization (CPPO) to address this issue. CPPO integrates sample-wise weighting into the Proximal Policy Optimization (PPO) algorithm, effectively balancing policy learning and knowledge retention. Zhang et al. [35] propose Continual Optimal Policy Regularization (COPR), which calculates the optimal policy distribution without the partition function and uses previous optimal policy to regularize the current policy . Unlike traditional methods, COPR operates within a single learning phase and incorporates a scoring module to facilitate learning from unlabeled data, enhancing its adaptability for continual learning without direct human feedback.

Sun et al. [36] introduce LAMOL, which generates pseudo-samples from previous tasks while training on a new task. It effectively mitigates knowledge loss without requiring additional memory or computational resources. Building on this framework, Wang et al. [37] developed RVAE_LAMOL, which integrates a residual variational autoencoder (RVAE) to encode input data into a unified semantic space, thereby enhancing task representation. This model also incorporates an identity task to enhance the model's discriminative ability for task identification. To enhance training efficacy, the Alternate Lag Training (ALT) is devised to segment the training process into multiple phases

To reduce hallucinations in specialized domains such as the Chinese legal domain, Zhang et al. [38] propose a novel domain adaptation framework, named Adapt-Retrieve-Revise (ARR). It consists of three steps: adapting a 7-billion-parameter language model for initial responses, retrieving corroborative evidence from an external knowledge base, and integrating these to refine the final response with GPT-4.

Gogoulou et al. [116] study the pros and cons of updating a language model when new data comes from new languages – the case of continual learning under language shift. They feed various languages into the model to examine the impact of pre-training sequence and linguistic characteristics on both forward and backward transfer effects across three distinct model sizes. A new continual learning (CL) problem, named Continual Knowledge Learning (CKL), is introduced by Jang et al. [40]. To assess CKL approaches, the authors establish a benchmark and metric measuring knowledge retention, updating, and acquisition. Their work stresses the importance of adapting LLMs to evolving linguistic landscapes, providing robust baseline methodologies for future research.

Continual Pre-training Methods LLMs have demonstrated remarkable proficiency in tackling open-domain tasks. However, their application in specific domains faces notable challenges, encompassing the lack of domain-specific knowledge, limited capacity to utilize such knowledge, and inadequate adaptation to domain-specific data formats. To address these issues, researchers have explored a novel approach known as continual pre-training, aiming to adapt LLMs to specific domains [41, 43, 117]. Among these studies, Cheng et al. [117] draw inspiration from human learning patterns to develop a novel method that transforms raw corpora into reading comprehension texts. Furthermore, they discover that while training directly on raw data enhances the model's domain knowledge, it significantly hurts the question-answering capability of the model.

Domain-adaptive Continual Pre-training (DACP) uses a large domain corpus, leading to high costs. To reduce these, Xie et al. [41] propose two strategies: Efficient Task-Similar Domain-Adaptive Continual Pre-training (ETS-DACP) and Efficient Task-Agnostic Domain-Adaptive Continual Pre-training (ETA-DACP). ETS-DACP is tailored to improve performance on specific tasks, building task-specific foundational LLMs. Conversely, ETA-DACP selects the most informative samples across the domain. Given the high cost of training LLMs from scratch and limited annotated data in certain domains, Ma et al. [43] propose a novel model called EcomGPT-CT. It employs a fusion strategy to exploit semi-structured E-commerce data. Moreover, multiple tasks are designed to assess LLMs' few-shot in-context learning ability and zero-shot performance after fine-tuning.

Parameter-Efficient Tuning Methods. Chen et al. [39] present an innovative Lifelong Learning framework, termed Lifelong-MoE, which leverages a Mixture-of-Experts (MoE) architecture (Figure 8). This architecture enhances the model's capacity by incorporating new experts. The previously trained experts and gating mechanisms are frozen to mitigate CF.

3. VLMs-based DIL

Vision-language models (VLMs) have demonstrated superiority in domain-incremental learning contexts. Yi et al.

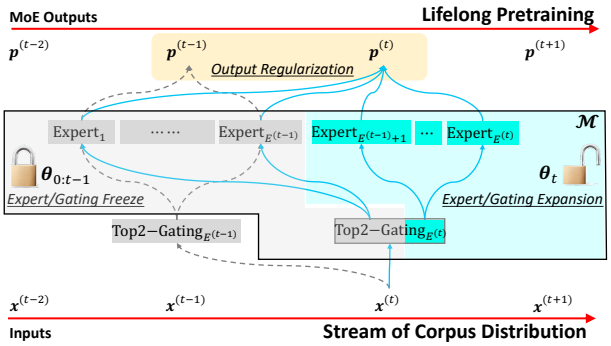

Fig. 8: The framework of Lifelong-MoE, taken from [39].

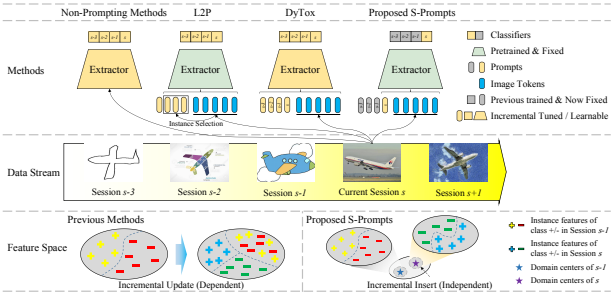

Fig. 9: The framework of S-Prompts, taken from [44].

[45] integrate VLMS with continual learning methodologies to develop a general-purpose medical AI. Moreover, their study highlights the significance of data-efficient adaptation algorithms in minimizing the necessity for extensive labeling when transitioning to new domains or tasks. Furthermore, prompt text is utilized to master tthe pre-trained knowledge embedded within VLMs. Aiming to independently learn prompts across disparate domains by using pre-trained visual-language models, S-Prompt [44] is devised (Figure 9). This method encompasses techniques for acquiring image prompts and introduces an innovative methodology for language-image prompt acquisition. Prompt learning is conducted separately, utilizing a unified cross-entropy loss function during training. During inference, a K-NN (k-nearest neighbors) technique is employed to discern the domain. This approach delineates distinct subspaces for each domain, thereby enhancing class separability, mitigating or obviating forgetting, and bolstering transfer learning capabilities.

In the domain of visual question answering, Zhang et al. [46] introduce VQACL, a novel framework designed to effectively integrate data from both visual and linguistic modalities. This integration is achieved through a dual-level task sequence that enhances the model's performance on complex multimodal tasks. Central to VQACL is a compositionality test that evaluates the model's ability to generalize new skill and concept combinations. The framework also incorporates a novel representation learning strategy that differentiates between sample-specific (SS) and sample-invariant (SI) features. SS features capture distinctive attributes of individual inputs, enhancing output uniqueness, while SI features, derived from category prototypes, ensure essential characteristics are retained. This approach mitigates catastrophic forgetting and enhances generalizability across tasks, improving the model's compositional abilities.

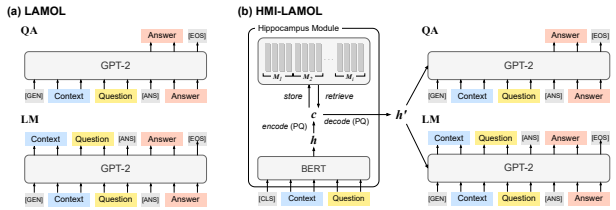

Fig. 10: The framework of HMI, taken from [55].

b. Task-Incremental Learning

1. PLMs-based TIL

Traditional Methods. Drawing inspiration from neurobiological mechanisms, Maekawa et al. [55] present an inventive approach known as Hippocampal Memory Indexing (HMI) to augment the generative replay technique. HMI leverages hippocampal memory indexing to integrate compressed representations of prior training instances, facilitating selective guidance for the generation of training samples. This methodological refinement contributes to heightened specificity, balance, and overall quality of the replayed samples. In tackling the Continual Few-Shot Relation Learning (CFRL) challenge, Qin et al. [51] propose ERDA as a solution, drawing upon replay- and regularization-based techniques. The ERDA framework integrates embedding space regularization and data augmentation strategies to effectively confront the task of acquiring new relational patterns from scarce labeled instances while mitigating the risk of catastrophic forgetting. Wang et al. [118] introduced a memory-based approach to continual learning, termed Episodic Memory Replay (EMR). This method leverages working memory by selectively replaying stored samples during each iteration of learning new tasks, thereby facilitating the integration of new knowledge while preserving previously acquired information.

Conure [58] is a framework that effectively manages multiple tasks by leveraging the redundancy of parameters in deep user representation models. Initially, it prunes less critical parameters to make room for new, task-specific ones. Subsequently, it incorporates these new parameters while retaining key parameters from prior tasks, facilitating positive transfer learning. To prevent the loss of previously acquired knowledge, it maintains these essential parameters in a fixed state.

Notably, TERACON [59] utilizes task-specific soft masks to isolate parameters, which not only targets parameter updates during training but also clarifies the relationships between tasks. This method includes a novel knowledge retention module that utilizes pseudo-labeling to mitigate the well-known problem of catastrophic forgetting.

Ke et al. [48] introduced a model known as CTR, which employs innovative techniques such as CL-plugin and task masking to tackle the issue of catastrophic forgetting and to facilitate knowledge transfer across tasks. These strategies are particularly effective when utilized in conjunction with pre-trained language models, such as BERT, enhancing their adaptability and efficacy. In the specific context of User Intent Classification (UIC) within large-scale industrial applications, Wang et al. [49] introduce a novel methodology, MeLL, which utilizes a BERT-based text encoder to gener-

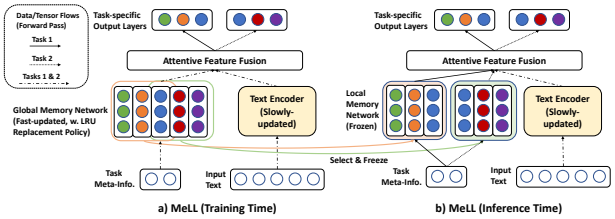

Fig. 11: The framework of MeLL, taken from [49].

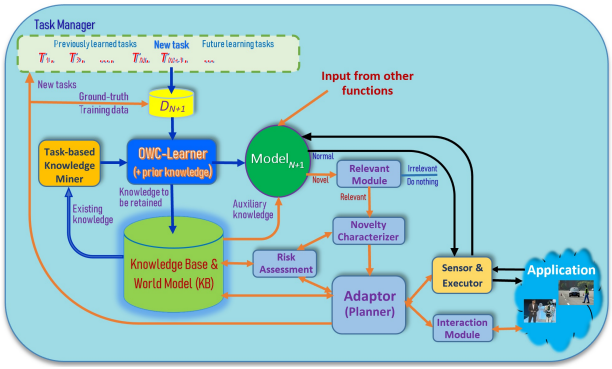

Fig. 12: The framework of LINC, taken from [119].

ate robust, dynamically updated text representations for continual learning. MeLL combines global and local memory networks to preserve prototype representations across tasks, acting as a meta-learner that rapidly adapts to new challenges. It employs a Least Recently Used (LRU) policy for efficient global memory management and minimizes parameter growth. Each UIC task benefits from a dedicated output layer equipped with an attentive summarization mechanism, optimizing the integration of diverse features to improve task performance.

Recent advancements in conversational AI have concentrated on mitigating the limitations of traditional chatbots, which are dependent on static knowledge bases and extensive manual data annotation. The Lifelong INteractive learning in Conversation (LINC) methodology represents a significant stride in this direction, embodying a dynamic learning framework that mimics human cognitive processes during interactions [50, 119, 120]. Structured around an Interaction Module, a Task Learner, and a Knowledge Store, LINC enables chatbots to dynamically integrate and utilize knowledge within conversational contexts (Figure 12). This approach not only facilitates real-time information extraction and learning from user interactions but also addresses challenges such as managing erroneous inputs, refining conversational skills, and maintaining user engagement, thereby enhancing the chatbots' linguistic and interactive capabilities.

Continual Pre-training Methods. Continual pre-training represents a paradigm wherein pre-trained language models (PLMs) are progressively enhanced through the assimilation of new knowledge from expanding datasets. Sun et al. [60] introduced a framework called ERNIE 2.0, which incrementally constructs pre-training tasks, enabling the acquisition of complex lexical, syntactic, and semantic nuances embedded within the training corpora. This model deviates from traditional fixed-task training, instead employing continual multi-task learning to refine its capabilities continually. Further advancing the field of continual pre-training, Recyclable-Tuning [61] introduces a novel concept of recyclable tuning that features two distinct strategies: initialization-based and distillation-based methods. The former uses fine-tuned weights of existing PLMs as the basis for further enhancements, capitalizing on the established parametric relationships. Conversely, the distillation-based approach harnesses outdated weights to maintain knowledge continuity and efficiency in successor models.

Parameter-Efficient Tuning Methods. Expounding upon the crucial need for more efficacious knowledge integration, Zhang et al. [56] propose a pioneering strategy which entails the integration of Adaptive Compositional Modules alongside a replay mechanism. These modules are designed to dynamically adjust to new tasks and are supplemented by pseudo-experience replay, significantly enhancing knowledge transfer. This framework is distinguished by its adaptive integration of modules within transformer architectures, skillfully orchestrating the interactions between existing and new modules to address emerging tasks. Additionally, the implementation of pseudo-experience replay promotes efficient knowledge transfer across these modules. Concurrently, Jin et al. [53] introduce the Continual Learning of Few-Shot Learners (CLIF) challenge, wherein a model accumulates knowledge continuously across a series of NLP tasks. Their investigation delves into the impact of continual learning algorithms on generalization capabilities and advances a novel approach for generating regularized adapters.

Instruction Tuning-based Methods. Zhao et al. [52] propose the Prompt Conditioned VAE for Lifelong Learning (PCLL) specifically designed for task-oriented dialogue (ToD) systems. PCLL employs a conditional variational autoencoder influenced by natural language prompts to generate high-quality pseudo samples, effectively capturing task-specific distributions. Additionally, a distillation process is integrated to refine past knowledge by reducing noise within pseudo samples. Razdaibiedina et al. [47] introduce Progressive Prompts (PP), a novel approach to continual learning in language models. PP addresses catastrophic forgetting without resorting to data replay or an excessive proliferation of task-specific parameters. This method involves acquiring a fresh soft prompt for each task, gradually appending it to previously learned prompts while keeping the base model unchanged. The ConTinTin paradigm [54] develops a computational framework for sequentially mastering a series of new tasks guided by explicit textual instructions. It synthesizes projected outcomes for new tasks based on instructions, while facilitating knowledge transfer from prior tasks (forward-transfer) and maintaining proficiency in previous tasks (backward-transfer).

In realm of lifelong in-context instruction learning aimed at enhancing the target PLM's instance- and task-level generalization performance as it observes more tasks, DYNAINST is devised by Mok et al. [57]. DYNAINST integrates the principles of parameter regularization and experience replay. The regularization technique employed by DYNAINST is tailored to foster broad local minima within the target PLM. In order to devise a memory- and computation-efficient experience replay mechanism, they introduce Dynamic Instruc-

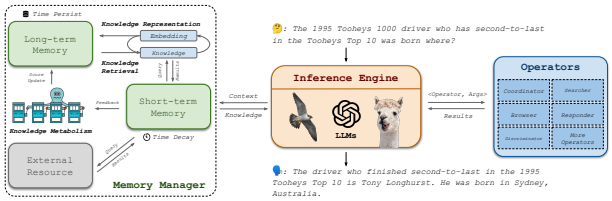

Fig. 13: The framework of DynaMind, taken from [65].

tion Replay, comprising Dynamic Instance Selection (DIS) and Dynamic Task Selection (DTS). DIS and DTS dynamically determine the selection of instances and tasks to be stored and replayed, respectively.

2. LLMs-based TIL

Recent attention has focused on the convergence of large language models (LLMs) with continual learning methodologies, exemplified by significant contributions such as those by Wang et al. [121] and Peng et al. [68]. Benefiting from vast corpora and advanced hardware infrastructure, LLMs showcase remarkable capabilities in language comprehension and generation. However, challenges arise in scenarios involving sequential tasks, where LLMs often exhibit a decline in performance known as catastrophic forgetting.

Traditional Methods. DynaMind, introduced by Du et al. [65], emerges as a pioneering framework which intricately incorporates memory mechanisms and modular operators, enhancing the precision of LLM outputs. Comprising three essential components, DynaMind includes a memory module dedicated to storing and updating acquired knowledge, a modular operator for processing input data, and a continual learning module responsible for dynamically adjusting LLM parameters in response to new knowledge. Furthermore, Luo et al. [122] conduct a thorough investigation into catastrophic forgetting (CF) in Large Language Models (LLMs) during continual fine-tuning. Their experiments across various domains, including domain knowledge, reasoning, and reading comprehension, reveal the prevalence of CF in LLMs ranging from 1b to 7b scale, with severity increasing with model size. Comparative analysis between decoder-only BLOOMZ and encoder-decoder mT0 indicates that BLOOMZ exhibits less forgetting. Additionally, LLMs demonstrate the ability to mitigate language bias during continual fine-tuning. Contrasting ALPACA against LLAMA, the study highlights ALPACA's superiority in preserving knowledge and capacity, suggesting that general instruction tuning aids in mitigating CF during subsequent fine-tuning phases. This research provides valuable insights into CF dynamics in LLMs, offering strategies for knowledge retention and bias mitigation.

Wang et al. [121] present TRACE, an innovative benchmark meticulously designed to evaluate continual learning capabilities in LLMs. Comprising eight distinct datasets spanning challenging tasks, including domain-specific challenges, multilingual capabilities, code generation, and mathematical reasoning, TRACE serves as a comprehensive evaluation platform. The authors rigorously examine the effectiveness of conventional Continual Learning (CL) methods when applied to LLMs within the TRACE framework.

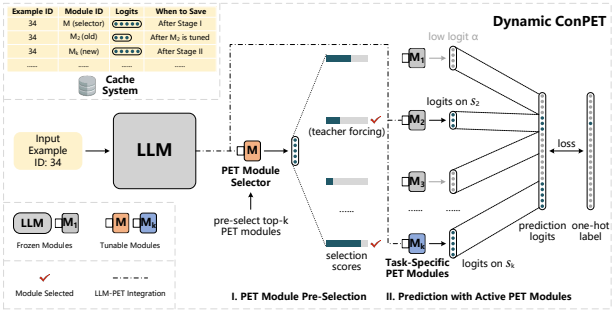

Fig. 14: The framework of ConPET, taken from [62].

Peng et al. [68] propose the Joint Adaptive ReParameterization (JARe) framework, enhanced with Dynamic Task-related Knowledge Retrieval (DTKR), to facilitate adaptive adjustment of language models tailored to specific downstream tasks. This innovative approach leverages task distribution within the vector space, aiming to streamline and optimize the continual learning process seamlessly.

Parameter-Efficient Tuning Methods. Large language models (LLMs) encounter several substantial challenges that limit their practical applications. These include high computational requirements, significant memory demands, and a tendency toward catastrophic forgetting. Such limitations highlight the need for ongoing research into more efficient and robust approaches to training and deploying these models. Continual Parameter-Efficient Tuning (ConPET) [62] is designed for the continuous adaptation of LLMs across diverse tasks, leveraging parameter-efficient tuning (PET) strategies to enhance both efficiency and performance. ConPET encompasses two primary modes: Static ConPET and Dynamic ConPET. Static ConPET adapts techniques previously utilized in smaller models for LLMs, thus minimizing tuning costs and reducing the risk of overfitting. Conversely, Dynamic ConPET enhances scalability by employing distinct PET modules for varying tasks, supplemented by a sophisticated selector mechanism.

Moreover, the ELM strategy [66] involves initially training a compact expert adapter on the LLM for each specific task, followed by deploying a retrieval method to select the most appropriate expert LLM for each new task. Furthermore, Wang et al. [67] have proposed orthogonal low-rank adaptation (O-LoRA), a straightforward yet efficacious method for facilitating continual learning in language models. O-LoRA mitigates catastrophic forgetting during task acquisition by employing distinct low-rank vector subspaces maintained orthogonally to minimize task interference. This method highlights the potential of orthogonal subspace techniques in improving the adaptability of language models to new tasks without compromising previously acquired knowledge.

Instruction Tuning-based Methods. Scialom et al. [64] introduced Continual-T0, an innovative framework aimed at exploring the capabilities of large language models (LLMs) through continual learning, incorporating rehearsal techniques. A central aspect of this approach is the employment of instruction tuning, a key strategy designed to enhance the adaptability and effectiveness of LLMs when encountering novel tasks. Leveraging self-supervised pre-training,

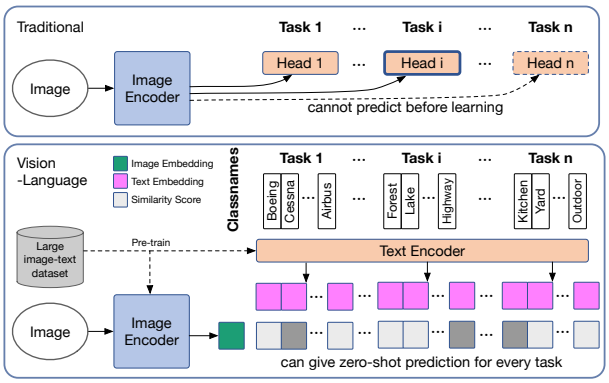

Fig. 15: The framework of ZSCL, taken from [70].

Continual-T0 demonstrates exceptional proficiency in mastering new language generation tasks while maintaining high performance across a diverse range of 70 previously encountered datasets. Despite the demonstrated proficiency of LLMs in adhering to instructions, their ability to generalize across underrepresented languages remains suboptimal. In response, InstructAlign [63] is proposed to address this challenge by aligning newly introduced languages with those previously learned, which possess abundant linguistic resources, thereby mitigating instances of catastrophic forgetting. The core novelty of this approach lies in its advancement of language adaptation methodologies for instruction-tuned LLMs, with particular emphasis on integrating underrepresented languages.

3. VLMs-based TIL

The long-term sustainability of pre-trained visual-language models (VLMs) is increasingly under scrutiny due to their dependence on continually expanding datasets. Although these models demonstrate robust performance across a diverse range of downstream tasks, the incessant growth of real-world data poses substantial challenges to the sustainability of traditional offline training methodologies.

Traditional Methods. CTP, presented by Zhu et al. [69] employs topology preservation and momentum contrast, which maintain consistent relationships within sample mini-batches across tasks, thereby preserving the distribution of prior embeddings. CTP also introduces the P9D dataset, comprising over one million image-text pairs across nine domains, aimed at visual language continuous pre-training (VLCP). Zheng et al. [70] address the issue of zero-shot transfer degradation in visual language models by introducing the Zero-Shot Continual Learning (ZSCL) method.This novel approach utilizes a label-free dataset to facilitate distillation in the feature space, coupled with the application of weight regularization within the parameter space. Furthermore, they introduce a new benchmark, the Multi-Domain Task Incremental Learning (MTIL), designed to evaluate incremental learning strategies across various domains, thereby enhancing the assessment of such methods. Moreover, ZSCL has also been adapted for use in the CIL setting, further broadening its applicability

Instruction Tuning-based Methods. By decoupling prompts and prompt interaction strategies, TRIPLET [72] effectively captures complex interactions between modalities. This includes specific designs for visual, textual, and fused prompts, as well as how to interact between different tasks through these prompts and retain crucial information, thereby reducing catastrophic forgetting. Decoupled prompts are designed to separate prompts in terms of multi-modality, layer-wise, and complementary, with each type of prompt containing learnable parameters intended to capture modality-specific knowledge from pre-trained visual-language models and training data. The prompt interaction strategies consist of three main components: the Query-and-Match Strategy, the Modality-Interaction Strategy, and the Task-Interaction Strategy. These components work together to enhance the model's adaptability to different tasks and its memory for old tasks.

Moreover, COIN [123] introduces a new continuous instruction tuning benchmark designed to evaluate the performance of Multimodal Large Language Models (MLLMs) in the sequential instruction fine-tuning paradigm. CoIN includes 10 commonly used data sets covering 8 task categories, ensuring the diversity of instructions and tasks. In addition, the trained model is evaluated from two aspects: instruction following and general knowledge, which respectively evaluate the consistency with human intention and the preservation of knowledge for reasoning. CoIN converts commonly used visual-linguistic datasets into instruction fine-tuning data formats by using different templates to explore the behavior of MLLMs in continuous instruction fine-tuning. This method takes into account the diversity between different tasks and attempts to enhance the model's adaptability to new and old tasks through diversified instruction templates. To alleviate the catastrophic forgetting problem of MLLMs, CoIN introduces the MoELoRA method. This method reduces forgetting by using different experts to learn knowledge on different tasks and using gate functions to regulate the output of these experts.

Parameter-Efficient Methods. MoE-Adapters4CL [71] introduces a parameter-efficient continual learning method to mitigate long-term forgetting in incremental learning of visual-language models. Their approach involves dynamically extending the pre-trained CLIP model to accommodate new tasks by integrating a Mixture-of-Experts (MoE) adapter. Specifically, MoE consists of several LoRA adapter experts and routers, where the router calculates gating weights and uses the $TopK$ function to select the k most relevant experts for learning the current task. To maintain the zero-shot recognition capabilities of the visual-language model, a Distribution Discriminative Automatic Selector (DDAS) is further introduced, which can automatically route in-distribution and out-of-distribution inputs to the MoE adapters and the original CLIP, respectively. Furthermore, the MoE-Adapters4CL framework has also been adapted for use in the CIL setting.

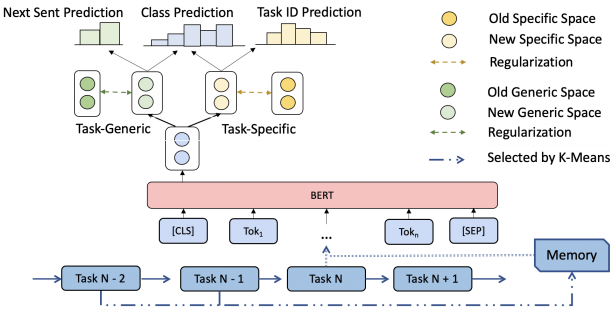

Fig. 16: The framework of IDBR, taken from [74].

c. Class-Incremental Learning

1. PLMs-based CIL

Traditional Methods. The study presented in [77] introduces ExtendNER, a novel framework for continual learning in Named Entity Recognition (NER) that obviates the need for extensive re-annotation. This framework employs a knowledge distillation (KD) technique, wherein a pre-existing named entity recognition (NER) model, termed the "teacher", imparts knowledge to a newly developed model, termed the "student". The student model, designed to identify new entity types, progressively learns by emulating the teacher model's responses to a new dataset. This method allows the student to acquire the ability to recognize new entities while retaining knowledge of previously identified ones. Additionally, Liu et al. [80] propose an innovative method for learning distributed representations of sentences. This method initiates with the configuration of sentence encoders using features independent of any specific corpus, followed by iterative updates through Boolean operations on conceptor matrices. This technique ensures that the encoders maintain their performance on existing datasets while adapting effectively to new data.

Huang et al. [74] introduced an innovative methodology known as Information Disentanglement-based Regularization (IDBR) to address the enduring challenges associated with continual text classification. This method effectively disentangles the hidden spaces of text into task-generic and task-specific representations, employing distinct regularization strategies to enhance knowledge retention and facilitate generalization. Furthermore, the integration of two auxiliary tasks, namely next sentence prediction and task-id prediction, serves to augment the learning process by reinforcing the separation between generic and specific representational spaces.

Instruction Tuning-based Methods. Introduced by Varshney et al. [75], Prompt Augmented Generative Replay (PAGeR) is a method that enables continual learning in intent detection without retaining past data. PAGeR leverages pre-trained language models to generate intent-specific utterances specific to new intents while preserving accuracy on existing ones. Unlike exemplar replay, which stores specific examples, PAGeR is structured to selectively maintain relevant contexts for each specified intent, which are then employed as generation prompts. This approach combines utterance generation and classification into one model, enhancing knowledge transfer and optimization.

Parameter-Efficient Tuning Methods. In the domain of task-oriented dialogue systems, Continual Few-shot Intent Detection (CFID) focuses on recognizing new intents with few examples. Li et al. [78] propose Prefix-guided Lightweight Encoder (PLE) to address this by using a parameter-efficient tuning method that combines a Continual Adapter module with a frozen Pre-trained Language Model (PLM) and a Prefix-guided Attention mechanism to reduce forgetting. To further mitigate forgetting, the Pseudo Samples Replay (PSR) strategy reinforces prior knowledge by replaying crucial samples from previous tasks. The Teacher Knowledge Transfer (TKT) strategy uses distillation to transfer task-specific knowledge to maintain performance on new tasks. Additionally, the Dynamic Weighting Replay (DWR) strategy dynamically adjusts weights of previous tasks to balance new knowledge acquisition with the revision of old tasks, navigating the variability and potential negative impacts of prior tasks

The Efficient Parameter Isolation (EPI) method, introduced in Wang et al. [73], assigns unique subsets of private parameters to each task alongside a shared pre-trained model. This approach ensures precise parameter retrieval and has been shown to outperform non-caching methods in continual language learning benchmarks, while remaining competitive with caching methods. Furthermore, EPI employs random static masking to reduce storage requirements, increasing its viability in resource-constrained environments.

In complex system environments, efficient and adaptable machine learning architectures are crucial, especially for classification tasks with sequentially presented data. Wójcik et al. [79] devise a novel architecture, Domain and Expertise Ensemble (DE&E), comprising a feature extractor, classifier, and gating mechanism. The feature extractor employs a multi-layer neural network to convert input data into embeddings, while the classifier, a mixture of binary class-specific experts, leverages a gating mechanism to select the appropriate expert for the current input dynamically. This Mixture of Experts-based method promotes incremental learning by training experts with class-specific samples and combines their outputs during testing to derive the final classification.

2. LLMs-based CIL

In [81], Cao et al. introduce a framework for a Generative Multi-modal Model (GMM) that leverages large language models for class-incremental learning. This innovative approach entails the generation of labels for images by employing an adapted generative model. Following the production of detailed textual descriptions, a text encoder is utilized to extract salient features from these descriptions. These extracted features are subsequently aligned with existing labels to ascertain the most fitting label for classification predictions.

3. VLMs-based CIL

Traditional Methods. The paper by Kim et al. [82] introduces VLM-PL, a novel approach for class incremental object detection that incorporates new object classes into a detection model without forgetting old ones. Utilizing a visual-language model (VLM), this method enhances the pseudo-

labeling process to improve the accuracy and performance of object detection in continual learning settings. VLM-PL starts with a pre-trained detector to generate initial pseudo-labels, which are then validated through a visual-language evaluation using a specially designed hint template. Accurate pseudo-labels are retained and combined with ground truth labels to train the model, ensuring it remains proficient in recognizing both new and previously learned categories.

PROOF, as introduced by Zhou et al. [84], develops a method to enhance model memory retention when adapting to downstream tasks. This method involves a projection technique that maps pre-trained features into a new feature space designed to preserve prior knowledge. Additionally, to effectively utilize cross-modal information, PROOF introduces a fusion module that employs an attention mechanism. This module adjusts both visual and textual features simultaneously, enabling the capture of semantic information with enhanced expressive power.

Recently, jha et al. [86] presents a new approach for adapting pre-trained vision-language models like CLIP to new tasks without forgetting previous knowledge. It employs a Variational Inference framework to probabilistically model the distribution of visual-guided text features, enhancing fine-tuning reliability by accounting for uncertainties in visual-textual interactions. Key to CLAP is the visual-guided attention (VGA) module, which aligns text and image features to prevent catastrophic forgetting. Additionally, CLAP includes lightweight, task-specific inference modules that learn unique stochastic factors for each task, allowing continuous adaptation and knowledge retention.

Parameter-Efficient Tuning Methods. Liu et al. [83] introduce Adaptation-CLIP, which employs three strategies for CLIP's continual learning: linear adapter, self-attention adapter, and prompt tuning. The first two strategies add a linear layer and a self-attention mechanism, respectively, after the image encoder while freezing the remaining architecture. The third, prompt tuning, integrates trained prompts into the text encoder to enhance task comprehension and splices these with prior prompts to maintain continuity. To prevent catastrophic forgetting, a parameter retention strategy preserves significantly altered parameters from M_{t-1} to M_t, ensuring stability and effective continual learning.

Instruction Tuning-based Methods. The paper by Khan et al. [85] introduce two notable advancements: an enhanced prompt pool key query mechanism and category-level language guidance. The key query mechanism uses CLS features to improve prompt selection, featuring key replacement with a fixed CLS tag and dynamic mapping to task-level language representations, thereby enhancing accuracy and robustness across different tasks. Meanwhile, category-level language guidance is implemented in the vision transformer to better align output features with category-specific language representations, significantly improving task handling and category differentiation, leading to improved model performance.

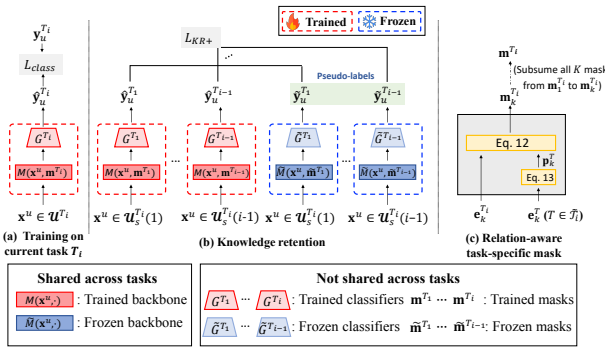

Fig. 17: The framework of TERACON, taken from [59].

V. Online Continual Learning

a. Hard Task Boundary

1. PLMs-based HTB

The Hard Task Boundary (HTB) setting has been developed to enable continuous knowledge acquisition by learning models from a dynamically changing stream of textual data, without the need for dataset identifiers. This approach has been applied to a range of downstream tasks across various domains, illustrating its broad applicability and effectiveness. For example, Shen et al. [91] have implemented HTB in slot filling, Michieli et al. [124] have utilized it in audio classification, and Vander et al. [125] have explored its use in automatic speech recognition.

Traditional Methods. Continual learning (CL) methodologies, particularly pertinent to online scenarios, encompass a variety of approaches. These include parameter-isolation-based methods [87, 88], replay-based methods [89] and regularization-based methods [125, 90].

MBPA++ [87] introduces a framework for lifelong language learning, enabling a pre-trained model to learn continually from textual examples without requiring labeled datasets. It employs an episodic memory system with sparse experience replay and local adaptation techniques to prevent catastrophic forgetting. Extending this framework, Meta-MBPA++ [88] integrates three core lifelong learning principles, enhancing performance in text classification and question answering tasks while using only 1% of the typical memory usage. Liu et al. [90] introduce a regularization-based strategy, CID, for lifelong intent detection. This method uses cosine normalization, hierarchical knowledge distillation, and inter-class margin loss to tackle the challenges of data imbalances in the lifelong intent detection (LID) task, aiming to mitigate the negative impacts associated with these imbalances.

b. Blurry Task Boundary

1. PLMs-based BTB

MBPA++ and Meta-MBPA++, as detailed in the works of [87] and [88] respectively, can also exemplify models capable of adapting to environments with indistinct task boundaries. TPEM [26] employs a tripartite approach within an encoder-decoder framework, consisting of pruning, expanding, and masking techniques. Pruning helps preserve essen-

tial information from previous tasks, expansion increases the model's capacity to accommodate new tasks, and masking reduces interference from previous tasks' weights, thus enhancing learning efficiency. Originally, *online meta-learning* (OML) [126] and *a neuromodulatory meta-learning algorithm* (ANML) [127] were intended to continuously learn new sequences of tasks during the testing phase. Holla et al. [89] adapt these algorithms for a conventional continual learning context, where the focus is on evaluating performance on previously encountered tasks. The enhanced versions, named OML-ER and ANML-ER, incorporate an episodic memory module designed for experience replay.

2. VLMs-based BTB

Traditional Methods. Cui et al. [94] propose the Dynamic Knowledge Rectification (DKR) framework, designed to mitigate the propagation of incorrect information in foundation language models. The DKR framework operates by initially leveraging an existing model to identify and exclude obsolete or erroneous knowledge when confronted with new data. Subsequently, a rectification process is employed to amend these inaccuracies while ensuring the preservation of valid data associations. This process is especially vital when integrating new data, as it prevents the perpetuation of outdated or incorrect information. In cases where data is inaccessible through existing models, DKR utilizes paired ground-truth labels to support the continuous evolution of knowledge bases, thereby enhancing the model's applicability and accuracy.

Parameter-Efficient Tuning Methods. Wang et al. [92] introduce the Continual Bias Adaptor (CBA), a novel method designed to enhance the efficacy of online continual learning (CL) by mitigating catastrophic forgetting. The CBA method dynamically modifies the classifier's network to adapt to significant shifts in data distribution observed during training, thereby preserving knowledge from earlier tasks. Notably, the CBA module can be deactivated during testing phases, eliminating additional computational burdens or memory demands. This feature highlights the practicality and applicability of the CBA in real-world scenarios.

Instruction Tuning-based Methods. In the BTB scenario, the Mask and Visual Prompt tuning (MVP) method, as detailed by Moon et al. [93], addresses challenges like intra- and inter-task forgetting and class imbalance effectively. MVP features instance-wise logit masking to prevent irrelevant information retention, contrastive visual prompt tuning loss to ensure consistent prompt selection, gradient similarity-based focal loss to focus on overlooked samples, and adaptive feature scaling to balance the integration of new knowledge with existing data retention. Collectively, these mechanisms enhance MVP's robustness in maintaining knowledge consistency and managing class imbalance in the Si-Blurry scenario.

Online Learning for CV Recent advancements in class-incremental online learning predominantly address computer vision tasks, encapsulating various methodologies such as regularization-based [128, 129, 130, 131], replay-based [132, 133, 134, 135, 136], distillation-based [137, 138, 139], and gradient-based approaches [140, 141, 142, 143, 144]. Among these, Koh et al. [145] introduced the Class-Incremental Blurry (CLIB) model, which distinguishes itself through its task-free nature, adaptability to class increments, and prompt response to inference queries, showing superior performance over traditional continual learning methods. In a related vein, Gunasekara et al. [146] explore the Online Streaming Continual Learning (OSCL), a hybrid of Stream Learning (SL) and Online Continual Learning (OCL), which integrates aspects of both domains. Furthermore, issues of shortcut learning and bias in online continual learning have been tackled by Wei et al. [147] and Chrysalis et al. [148]. Additionally, Semola et al. [149] propose Continual-Learning-as-a-Service (CLaaS), a service model that leverages continual learning to monitor shifts in data distribution and update models efficiently. This array of developments highlights the dynamic capabilities of online continual learning frameworks to effectively address complex challenges across a variety of computer vision tasks.

In this section, we review the typical datasets and metrics for continual learning.

VI. DATASETS

a. Offline Datasets for NLP

1. Datasets for Classification.

Text Classification. The most typical task for continual learning is text classification. The foundational text classification benchmark encompasses five text classification datasets introduced by [155], including AG News, Amazon Reviews, Yelp Reviews, DBpedia, and Yahoo Answers [36]. Particularly, the AG News dataset has 4 classes for news classification; the Amazon and Yelp dataset has 5 classes for sentiment analysis; the DBpedia dataset has 14 classes for Wikipedia text classification; and the Yahoo dataset has 10 classes for Q&A classification. The text classification benchmark includes 115,000 training and 7,600 test examples for each task, holding out 500 samples per class from the training set for validation.

Building upon this, Razdaibiedina et al. [47] developed a novel continual learning (CL) benchmark. This benchmark not only utilizes the foundational text classification benchmark but also integrates additional datasets from the GLUE benchmark [156], SuperGLUE benchmark [157], and the IMDB dataset [158]. Specifically, the GLUE benchmark datasets included are MNLI, QQP, RTE, and SST2, focusing on tasks like natural language inference, paraphrase detection, and sentiment analysis. Similarly, the SuperGLUE datasets—WiC, CB, COPA, MultiRC, and BoolQ—encompass tasks ranging from word sense disambiguation to question answering. Each of these benchmarks includes three different dataset versions with 20, 200, and 1000 training samples per class, respectively, and test set performance is reported for each configuration. Additionally, 500 samples per class are designated for validation to facilitate cross-dataset comparisons and evaluate model generalizability across diverse data types and sizes. DE&E [79] uses three common text classification data sets with differ-

Datasets	#Train	#Val	#Test	#Total	CL Settings	NLP Problems	Language	#D/T/C
					Offline			
Progressive Prompts [47]	-	-	-	-	TIL	Sentiment analysis, topic classification, boolean QA, QA, paraphrase detection, word sense disambiguation, natural language inference	English	15 tasks
MeLL [49]	1,430,880	173,781	118,240	1,722,901	TIL	Intent classification	English	1184 tasks
Continual-T0 [97]	800,000	-	33,382	833,382	TIL	Text Simplification, Headline Generation with Constraint, Haiku Generation, Covid QA, Inquisitive Question Generation, Empathetic Dialogue Generation, Explanation Generation, Twitter Stylometry	English	8 tasks
COPR [35]	-	-	-	-	TIL	QA tasks, summary task, positive file review generation task	English	3 tasks
Adaptive Compositional Modules [56]	50,725	-	27,944	78,669	TIL	Natural language generation, SQL query generation, Summarization and Task-oriented dialogue	English	8 tasks
CODETASKCL [150]	181,000	9,700	10,000	200,700	TIL	Code generation, code translation, code summarization, and code refinement	Hybrid	4 tasks
Lifelong Simple Questions [118]	-	-	-	-	TIL	single-relation questions	English	20 tasks
Lifelong FewRel [118]	-	-	-	-	TIL	few-shot relation detection	English	10 tasks
InstrDialog [151]	9,500	950	1,900	12,350	TIL	Dialogue state tracking, dialogue generation, intent identification	English	19 tasks
InstrDialog++ [151]	3,800	1,900	3,800	9,500	TIL	Dialogue Generation, Intent Identification, Dialogue State Tracking, Style Transfer, Sentence Ordering, Word Semantics, Text Categorization, Pos Tagging, Fill in The Blank, Program Execution, Question Generation, Misc, Coherence Classification, Question Answering, Summarization, Commonsense Classification, Wrong Candidate Generation and Toxic Language Detection	English	38 tasks
ConTinTin [54]	-	-	-	-	TIL	Question generation tasks (QG), answer generation tasks (AG), classification tasks (CF), incorrect answer generation tasks (IAG), minimal modification tasks (MM) and verification tasks (VF)	English	61 tasks
Tencent TL [58]	-	-	-	-	TIL	personalized recommendations and profile predictions	English	6 tasks
Movielens [58]	-	-	-	-	TIL	personalized recommendations and profile predictions	English	3 tasks
NAVER Shopping [59]	-	-	-	-	TIL	search query prediction tasks	English	6 tasks
TRACE [121]	40,000	-	16,000	56,000	TIL	Domain-specific task, Multi-lingual task, Code completion task, Mathematical reasoning task	Hybrid	8 tasks
ABSC [22]	3,452	150	1,120	4,722	DIL	Aspect-based sentiment classification	English	19 Domains
DecaNLP [36]	169,824	-	32,116	201,940	DIL	Question answering, semantic parsing, sentiment analysis, semantic role labeling, and goal-oriented dialogue	English	5 domains
Foundational text classification [36]	115,000	-	7,600	122,600	DIL	News classification, sentiment analysis, Wikipedia article classification, and question-and-answer categorization	English	5 domains
RVAE_LAMOL [37]	15,870	-	5,668	21,538	DIL	Oriented dialogue of the restaurant reservation task, semantic role labeling, sentiment classification	English	3 domains
COPR [35]	-	-	-	-	DIL	-	English	18 domains
SGD [29]	38,745	5,210	11,349	40,287	DIL	Dialogue state tracking	English	19 Domains
CPT [28]	3,121,926	-	-	3,121,926	DIL	Domain-adaptive pre-training task	English	4 Domains
CKL[40]	-	-	-	30,372	DIL	Domain-adaptive pre-training task	English	3 Domains
ELLE [23]	-	-	-	-	DIL	Domain-adaptive pre-training task	English	5 Domains
Domain-incremental Paper Stream [152]	-	-	-	-	DIL	Relation extraction and named entity recognition	English	4 domains
Chronologically-ordered Tweet Stream [152]	-	-	-	-	DIL	multi-label hashtag prediction and single-label emoji prediction	English	4 domains
AdapterCL [24]	31,426	4,043	4,818	40,287	DIL	Intent classification, Dialogue State Tracking (DST), Natural Language Generation (NLG), end-to-end (E2E) modeling	English	37 Domains
DE&E [79]	28,982	-	12,089	41,071	CIL	Text classification	English	3 tasks
EPI [73]	12,840	3,524	6,917	23,281	CIL	Text classification, topic classification	English	13 Classes
PAGeR [75]	59,754	7,115	15,304	82,173	CIL	Intent classification	English	355 Classes
PLE [78]	4,669	4,650	31,642	40,961	CIL	Intent classification	English	477 Classes
CoNLL-03 [77]	23,326	5,902	5,613	34,841	CIL	Named Entity Recognition	English	4 Classes
OntoNotes [77]	107,169	16,815	10,464	134,448	CIL	Named Entity Recognition	English	6 Classes
					Online			
Foundational text classification [87]	115,000	-	7,600	122,600	Hard and Blurry	News classification, sentiment analysis, Wikipedia article classification, and question-and-answer categorization	English	5 tasks
MBPA++ [87]	881,000	35,000	38,000	954,000	Hard and Blurry	News classification, sentiment analysis, Wikipedia article classification, questions and answers categorization, question answering	English	9 tasks
Lifelong FewRel [89]	-	-	-	-	Hard and Blurry	few-shot relation detection	English	10 tasks
Firehose[153]	-	-	-	110,000,000	Blurry	Personalized online language learning	English	1 tasks
TemporalWiki [154]	-	-	-	-	-	-	English	-

TABLE 1: THE STATISTICS INFORMATION OF THE EXISTING CL DATASETS. #D/T/C MEANS THE NUMBER OF DOMAINS/TASKS/CLASSES FOR DIL, TIL AND CIL, RESPECTIVELY.

ent characteristics-News-groups, BBC News, and Consumer Finance Complaints2. Such dataset can be used to evaluate model on tasks with with different difficulty levels.

The datasets introduced in [73] further are categorized into two groups based on the domain relevance between tasks: far-domain and near-domain. The far-domain group comprises two text classification tasks, which are foundational benchmarks [155] divided into topic classification (AG

News, Yahoo Answers, DBpedia) and sentiment classification (Yelp, Amazon Reviews). In contrast, the near-domain group uses the Web of Science (WOS) [159] and 20 Newsgroups [160], which are restructured according to their high inter-task relevance. The WOS dataset comprises seven parent classes, each with five closely related sub-classes, while the 20 Newsgroups dataset, containing six news topics, is reorganized into four tasks to maximize inter-task correlation. This bifurcation facilitates a nuanced assessment of continual learning methodologies based on task and domain relevancies.

Intent Classification. Some studies focus on intent classification tasks, where the classes are quite different in various domains or scenarios. In the realm of intent classification and detection, several datasets have been specifically designed to advance the field by addressing different challenges and providing diverse environments for model training and evaluation. The dataset, as introduced in PAGeR [75], aims to tackle the lifelong intent detection problem by combining three public intent classification datasets (CLINC150 [161], HWU64 [162], BANKING77 [163]), one text classification dataset (Stackoverflow S20 [164]), and two public multidomain dialog intent detection datasets (SGD [165], MWOZ [166]). Moreover, FewRel [167] is also incorporated to tackle the lifelong relation extraction problem. This integration is intended to simulate real-world applications by encompassing a broad spectrum of domains and query distributions, thereby facilitating the development of more robust and versatile intent detection systems.

Conversely, the dataset compiled in PLE [78] consolidates nine well-regarded intent detection datasets, including CLINC150 [161] and HWU64 [162], among others, arranged in a fixed random sequence to form a standardized benchmark. This dataset emphasizes the importance of consistency and comparability in performance evaluations across different intent detection models, providing a platform for assessing and enhancing various methodologies.

The dataset described by MeLL [49] specifically addresses intent detection within two distinct contexts: task-oriented dialogues (TaskDialog-EUIC) and real-world e-commerce interactions (Hotline-EUIC). TaskDialog-EUIC integrates data from Snips [168], TOP semantic parsing [169], and Facebook's Multilingual Task Oriented Dataset [170] into 90 tasks with overlapping label sets, amounting to over ten thousand samples. Hotline-EUIC is derived from an e-commerce dialogue system [171] and the hotline audios are transcribed to text by a high-accuracy industrial Automatic Speech Recognition (ASR) system. Both datasets designate 30% of their tasks as "base tasks" for initial training and use the rest for lifelong learning, with data randomly divided into training, development, and testing sets for each task.

Fine-grained Sentiment Analysis. Ke et al. [22] developed a dataset specifically tailored for task incremental learning in the context of aspect-based sentiment classification (ABSC). This dataset aggregates reviews from four distinct sources, thereby enhancing its diversity and applicability across multiple domains. The sources include the L5Domains dataset by Hu et al. [172], which features consumer reviews for five different products; the Liu3Domains dataset by Liu [173], comprising reviews pertaining to three products; the Ding9Domains dataset by Ding et al. [174], which includes reviews of nine varied products; and the SemEval14 dataset, which is focused on reviews of two specific products—laptops and restaurants.

2. Datasets for Generation.

In the rapidly advancing field of machine learning, diverse datasets function as crucial benchmarks for exploring various dimensions of language and code generation. These datasets address both universal and task-specific challenges, enabling a comprehensive evaluation of model capabilities. A particularly significant dataset highlighted in the work by Continual-T0 [97] focuses on English language generation tasks, including text simplification and empathetic dialogue generation, among others [175, 176]. The design of this dataset maintains uniformity in size, facilitating effective comparative analyses of performance across distinct tasks by ensuring a consistent volume of data for training. In a subsequent study, Luo et al. [122] conduct an analysis of catastrophic forgetting on Bloomz [177] using Continual T0 datasets.

The dataset, introduced in LAMOL [36], integrates elements from both DecaNLP [178] and the foundational text classification benchmark [155]. This dataset encompasses five distinct NLP tasks originally sourced from DecaNLP: question answering, semantic parsing, sentiment analysis, semantic role labeling, and goal-oriented dialogue. For the purposes of this dataset, all tasks, whether derived from DecaNLP or the foundational text classification benchmark, are restructured into a uniform format, conceptualized under the framework of a question answering (QA) task. Moreover, the dataset devised in RVAE_LAMOL [37], employs three tasks from DecaNLP: the English Wizard of Oz (WOZ) for goal-oriented dialogue on restaurant reservations, QA-SRL for semantic role labeling in a SQuAD-style format, and SST, which is a binary version of the Stanford Sentiment Treebank categorizing sentiments as positive or negative. These tasks are specifically treated as sequence generation tasks.

The dataset introduced in COPR [35] represents a pioneering effort in applying both Task Incremental Learning (TIL) and Domain Incremental Learning (DIL) within the context of benchmarks that utilize existing human preferences. Specifically, the TIL framework in this dataset mandates that the model sequentially acquires knowledge from three distinct tasks. These include the question-answering task utilizing the HH-RLHF dataset [179], the summarization task based on the Reddit TL, DR dataset with human feedback [180], and the positive film review generation task using the IMDB dataset [158]. Meanwhile, the DIL framework requires the model to adapt to three distinct segments from the SHP dataset, as described by Ethayarajh et al. [181]. This dataset comprises eighteen domains, systematically categorized into three groups based on the most significant observed performance degradation.

The dataset described in Adaptive Compositional Modules [56] explores sequence generation and categorizes tasks into "similar" and "dissimilar" groups based on their characteristics. Tasks classified as similar, including E2ENLG [182] and four domains (restaurant, hotel, TV, laptop) from

RNNLG [183], demonstrate shared patterns and are tested across four sequence orders, comprising a total of five tasks. In contrast, dissimilar tasks such as WikiSQL (SQL query generation) [184], CNN/DailyMail (news article summarization) [185], and MultiWOZ (semantic state sequence generation) [166] exhibit significant distributional shifts from previously encountered tasks. This categorization is critical as it poses unique challenges in terms of knowledge retention and adaptability, thereby allowing us to rigorously evaluate the model's capabilities in managing domain variability and knowledge across a diverse set of tasks.

The CODETASKCL dataset, explored by Yadav et al. [150], encompasses a diverse array of code-centric tasks, including code generation [186], summarization [187], translation [188], and refinement [189] across various programming languages. This dataset significantly enhances the breadth of language processing applications within technical fields. Furthermore, it highlights the practical utility of generating and refining code based on natural language inputs and existing code repositories. Such functionalities illustrate the interdisciplinary nature of machine learning applications in language processing.

3. Datasets for Information Extraction.

In the realm of natural language processing (NLP), various datasets are tailored to specific aspects of the task under continual learning paradigms. The dataset introduced in ExtendNER [77], exemplifies a continual learning approach to Named Entity Recognition (NER). This dataset amalgamates the CoNLL-03 English NER [190] and OntoNotes [191], covering a broad spectrum of entity types and sources. CoNLL-03, with its focus on four primary entity types (Person, Location, Organization, Miscellaneous) derived from news stories, contrasts with OntoNotes which spans a wider array of texts (e.g., news, conversations, weblogs) and annotates for a richer set of eighteen entity types, although only six (Organization, Person, Geo-Political Entity, Date, Cardinal, Nationalities and Religious Political Group) are prioritized for robust training samples. This hybrid dataset is structured to challenge the adaptability and generalization capabilities of NER systems across varied contexts.

Unlike the static nature of text in NER tasks, the Schema-Guided Dialog (SGD) [165] dataset, utilized in C-PT [29], serves the Dialog State Tracking aspect of IE, which involves maintaining the context of a dialog over time. The SGD dataset features 44 services across 19 domains, each treated as a separate task, and is designed to evaluate models on their ability to manage and extract information across conversational turns.

Lastly, the lifelong SimpleQuestions and lifelong FewRel datasets, devised in [118] is crafted for the task of relation extraction. It merges elements from the SimpleQuestions [192] and FewRel [167] to form a lifelong learning benchmark that confronts the challenges of relation detection in a few-shot context. This dataset is structured to assess the model's capability to incrementally learn and classify relations from vast vocabularies without losing prior knowledge. Furthermore, the incorporation of the k-means clustering technique plays a crucial role in the dataset. This clustering method systematically partitions relations into distinct clusters, which not only facilitates the incremental learning process but also ensures sustained efficiency and effectiveness as the model continuously adapts to new relation types.

4. Datasets for Continual Pre-training.

In the realm of continual pre-training for large language models (LMs), the development and utilization of specialized benchmarks play a pivotal role in evaluating and enhancing the effectiveness of continual learning systems. The dataset, introduced in CPT [28], primarily focuses on the continual post-training of LMs across a series of domain-specific, unlabeled datasets. It provides a rigorous test environment by using diverse corpora such as Yelp Restaurant Reviews [193], AI and ACL Papers [194], and AGNews articles [155]. Its main objective is to gauge how well an LM can incrementally integrate domain-specific knowledge without forgetting previously learned information, thereby enhancing its few-shot learning capabilities in these domains.

Contrary to the datasets employed in CPT [28], which evaluate domain-specific adaptability and incremental learning, the CKL benchmark [40] is meticulously designed to measure the LM's ability to retain timeless knowledge, update obsolete information, and acquire new knowledge. It comprises subsets like INVARIANTLAMA, UPDATED-LAMA, and NEWLAMA, which are crafted to probe specific types of knowledge that an LM may encounter in its learning trajectory. The construction of these subsets involves sophisticated methodologies including crowd-sourcing and expert validation, ensuring the relevance and reliability of the tasks designed to evaluate memory retention and learning efficacy during continual learning processes.

Whereas the aforementioned two datasets assess more controlled dimensions of knowledge integration and retention, the dataset introduced in ELLE [23] focuses on the dynamic scenario of accumulating streaming data from diverse sources in a lifelong learning context. This dataset mirrors the real-world challenge of a language model (LM) that must continuously adapt to new data inflows from multiple domains, including BOOKCORPUS (WB) [195], NEWS ARTICLES (NS) [196], AMAZON REVIEWS (REV) [197], BIOMEDICAL PAPERS (BIO) [194] and COMPUTER SCIENCE PAPERS (CS) [194]. The benchmark evaluates the LM's capacity to effectively integrate new information from these varied sources over time, highlighting the essential need for LMs to evolve in response to continual data growth and shifts in data distribution.

Jin et al. [152] constructs data streams to represent two prevalent types of domain shifts observed in practical scenarios. The first, a Domain-incremental Paper Stream, simulates the sequential evolution of research areas within academic papers, encompassing diverse disciplines such as biomedical and computer science. The second, a Chronologically-ordered Tweet Stream, models the temporal progression of tweets over time.

5. Datasets for Hybrid Tasks.

An increasing number of datasets are adopting a hybrid task approach that integrates multiple learning paradigms and task types, aimed at testing and enhancing the adaptabil-

ity of models. A notable example is the dataset introduced in AdapterCL [24], which is tailored for task-oriented dialogue systems. This dataset incorporates four task-oriented datasets: TaskMaster 2019 (TM19) [198], TaskMaster 2020 (TM20) [198], Schema Guided Dialogue (SGD) [165], and MultiWoZ [166]. These datasets have been pre-processed to form a curriculum encompassing 37 domains, structured under four continual learning settings: INTENT classification, Dialogue State Tracking (DST), Natural Language Generation (NLG), and end-to-end (E2E) modeling. This curriculum is pioneering in its breadth, offering a scope significantly wider than its predecessors, thus serving as an invaluable resource for examining the robustness of dialogue systems against domain shifts.

CITB [151], or Continual Instruction Tuning Benchmark, extends the concept of continual learning by focusing on instruction-based NLP tasks. Built on the comprehensive SuperNI [199] dataset, it includes over 1,600 tasks across diverse NLP categories. CITB differentiates itself by formulating two distinct streams—InstrDialog and InstrDialog++—to examine how models integrate and retain new dialogue-oriented and varied NLP tasks under continual learning settings. This benchmark suite not only tests task retention and adaptability but also explores how instruction tuning can be optimized for a continual learning framework.

The dataset introduced in ConTinTin [54] is an adaptation of the NATURAL-INSTRUCTIONS [**?**] dataset, specifically restructured to facilitate a continual learning framework. This adaptation involves decomposing the original crowdsourcing instructions into smaller, distinct sub-tasks to create a new dataset. Additionally, the new dataset incorporates an novel experimental design where tasks are selected randomly to create diverse sequences, enabling the evaluation of a model's adaptability to novel instructions without prior exposure. This benchmark highlights the significance of task diversity and dynamic sequence arrangements in continual learning, distinctively analyzing how the order of tasks impacts learning efficacy.

The dataset, used in Conure [58], consists of Tencent TL (TTL) [200] and Movielens (ML). The TTL dataset is designed to address three item recommendation tasks and three user profiling tasks, whereas the ML dataset exclusively focuses on three item recommendation tasks. Both datasets have been pre-processed to facilitate a continual learning framework, simulating environments where models must adapt to evolving data streams. Furthermore, Kim et al. [59] introduced the proprietary NAVER Shopping dataset, which builds upon the previously mentioned datasets. The NAVER Shopping dataset features six tasks: two for search query prediction, two for purchased item category prediction, and two for user profiling, all designed to meet real-world industry requirements.

Finally, the TRACE dataset, introduced by Wang et al. [121], is specifically designed to bridge the existing gap in the evaluation of large language models (LLMs) within the continual learning framework, encompassing a wide range of complex and specialized tasks. Distinguished from other datasets, TRACE targets domain-specific tasks that are multilingual and technical, including code completion and mathematical reasoning. This diversity presents a unique set of challenges that span both specialized and broad dimensions. Moreover, TRACE rigorously assesses the models' ability to sustain performance across tasks that demand different knowledge bases and cognitive skills. This evaluation highlights the essential need for adaptability in LLMs trained to operate under continual learning conditions, underscoring their potential in dynamic real-world applications.

b. Online Datasets for NLP

1. Datasets for Classification.

The foundational text classification benchmark, as introduced by Zhang et al. [155] has traditionally been applied in offline continual learning settings. Recent advancements have adapted this benchmark for online continual learning, notably in studies such as MBPA++ [87] and OML-ER [89]

2. Datasets for Generation.

The dataset, used in MBPA++ [87], comprises three distinct question-answering collections: SQuAD 1.1 [201], TriviaQA [202], and QuAC [203]. SQuAD 1.1 is a reading comprehension dataset based on Wikipedia articles, designed to assess the ability to derive answers from structured text. TriviaQA consists of question-answer pairs developed by trivia enthusiasts, accompanied by corroborative evidence sourced from both the web and Wikipedia, testing the model's capability to handle diverse information sources. QuAC adopts a dialog-style format in which a student queries about information in a Wikipedia article and a teacher responds using text directly from the article, challenging the model's interactive response generation. The distinctive characteristics inherent to each sub-dataset are crucial in establishing this dataset's significance for assessing the model's effectiveness across diverse question-answering scenarios.

3. Datasets for Information Extraction.

The lifelong relation extraction benchmark, used in OML-ER [89], is structed by Wang et al. [118] based on FewRel. Unlike the original application by Wang et al., the benchmark in OML-ER is adapted for online continuous learning scenarios.

4. Datasets for Other Tasks.

Hu et al. [153] compiled the Firehose dataset, consisting of 110 million tweets from over 920,000 users between January 2013 and September 2019. This dataset is split into FIREHOSE 10M and FIREHOSE 100M. The validation and testing involve three randomly sampled tweets per user, enabling the comparison of continual learning with traditional offline training methods and assessing temporal robustness.

Language models (LMs) are susceptible to becoming outdated due to ongoing changes in the world, which may lead to difficulties when tasked with processing information that has become relevant after their training period. This challenge is referred to as temporal misalignment. TemporalWiki [154] addresses this issue by serving as a lifelong benchmark that trains and evaluates LMs using consecutive snapshots of Wikipedia and Wikidata. This methodology assists in assessing an LM's capacity to both retain previously acquired knowledge and assimilate new information over time.

c. Offline CL Datasets for Multi-modal Tasks

The P9D dataset, introduced by Zhu et al. [69], consists of over one million image-text pairs from e-commerce data, organized into nine industry sector-based training tasks. It includes 1,014,599 training pairs, 2,846 for cross-modal retrieval tests, and 4,615 query pairs with 46,855 gallery pairs for multi-modal retrieval. Designed to mimic real-world e-commerce complexities, such as diverse backgrounds and multiple product views, P9D covers over 3,800 product classes, enhancing the challenge of feature clustering in multi-modal retrieval.

Qian et al. [72] introduce two novel benchmarks for continual learning, namely CL-TDIUC and CL-VQA2.0, which are derived from the TDIUC [204] and VQA2.0 [205], respectively. These benchmarks are categorized into three scenarios: the Continual Vision Scenario (ConVS), which deals with new visual scenes; the Continual Language Scenario (ConLS), focusing on new questions in existing scenes; and the Continual Vision-Language Scenario (ConVLS), addressing changes in both questions and visuals. ConVS and ConLS segment data by image and question hyper-categories, respectively, while ConVLS combines question types across image hyper-categories into five unique tasks to test adaptability across both modalities.

The dataset, introduced in DKR [94], comprises five benchmark datasets: MS-COCO Caption (MS-COCO) [206], Flickr30K [207], IAPR TC-12 [208], ECommerce-T2I (EC) [209], and RSICD [210]. Furthermore, two experimental scenarios are established. The first scenario involves a sequential processing of the datasets, specifically MS-COCO, Flickr30K, IAPR TC-12, EC, and RSICD, in that order. The second scenario, which builds on the approach proposed by Ni et al. [211], partitions the EC dataset into five sub-datasets for the training phase. The model's performance is subsequently tested on the Flickr30K, MS-COCO, and EC datasets.

d. Offline and Online Datasets for Other Tasks

In this paper, we provide a comprehensive review of recent studies in natural language processing (NLP) and multi-modal tasks, with a particular focus on the use of PLMs, LLMs and VLMs. Additionally, we introduce a range of offline continual learning (CL) datasets employed across various applications, including automatic speech recognition [125, 212], autonomous driving [213], disease classification [214], reinforcement learning [215], graph data [216], and computer vision [217, 218]. Furthermore, some online CL benchmarks are also proposed for computer version tasks, such as continual visual learning [219], continual object detection [220].

VII. METRICS

In this section, we review the principal metrics commonly used to evaluate continual learning. These metrics can be categorized into three main types: (1) overall performance, which assesses the algorithm's effectiveness across all tasks; (2) memory stability, which measures the extent to which an algorithm retains previously acquired knowledge; and (3) learning plasticity, which evaluates the algorithm's capacity to acquire new skills or knowledge. Each of these metrics provides insights into different aspects of the algorithm's performance in a continual learning context.

To begin, we establish the notation (Figure 18) used throughout the learning and evaluation phases of the model. Once the model completes a learning task, denoted as T_i, it evaluates its performance on a test set that encompasses all N tasks, where N is the total number of tasks in the set T. This evaluation is represented by a matrix $R \in \mathbb{R}^{N\times N}$, wherein each element $R_{i,j}$ indicates the model's test classification accuracy on task T_j after processing the final sample from task T_i.

a. Overall Performance.

The metric termed "Last" [221, 70] evaluates the overall performance of a continual learning (CL) method upon the completion of all tasks. Specifically, it computes the average score from the last row in the performance matrix R.

$$Last = \frac{1}{N}\sum_{i=1}^{N} R_{N,i} \tag{1}$$

Also, Zheng et al. [70] devise the "Avg" score metric, which computes the mean accuracy across all datasets and timestamps.

$$Avg = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} R_{i,j}\right) \tag{2}$$

In the seminal works of Rebuffi et al. [222] and Douillard et al. [223], the concept of Average Incremental Accuracy (AIA) is introduced. This metric is specifically designed to quantify the historical performance across different tasks. It calculates the average performance for each task by considering the lower triangular portion of the matrix R, effectively capturing the evolving competence of the system as new tasks are learned.

$$AIA = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{i}\sum_{j=1}^{i} R_{i,j}\right) \tag{3}$$

The metric, termed Transfer, is derived by computing the average of the performance values for tasks that are represented in the upper-right triangle of matrix R. This approach uniformly weights each dataset by averaging their performance across different tasks, thereby assessing the preservation of zero-shot transfer capabilities. Prior to commencing learning on task T_i, no fine-tuning is performed on tasks that precede T_i.

$$Transfer = \frac{1}{N-1}\sum_{i=2}^{N}\left(\frac{1}{i-1}\sum_{j=1}^{i-1} R_{j,i}\right) \tag{4}$$

Chaudhry et al. [224] devise a metric known as Learning Curve Area (LCA), which quantifies the speed of learning in a model. It first defines an average b-shot performance, where b represents the number of mini-batches, subsequent to the completion of training across all T tasks as follows:

$$Z_b = \frac{1}{N}\sum_{i=1}^{N} R_{N,i} \tag{5}$$

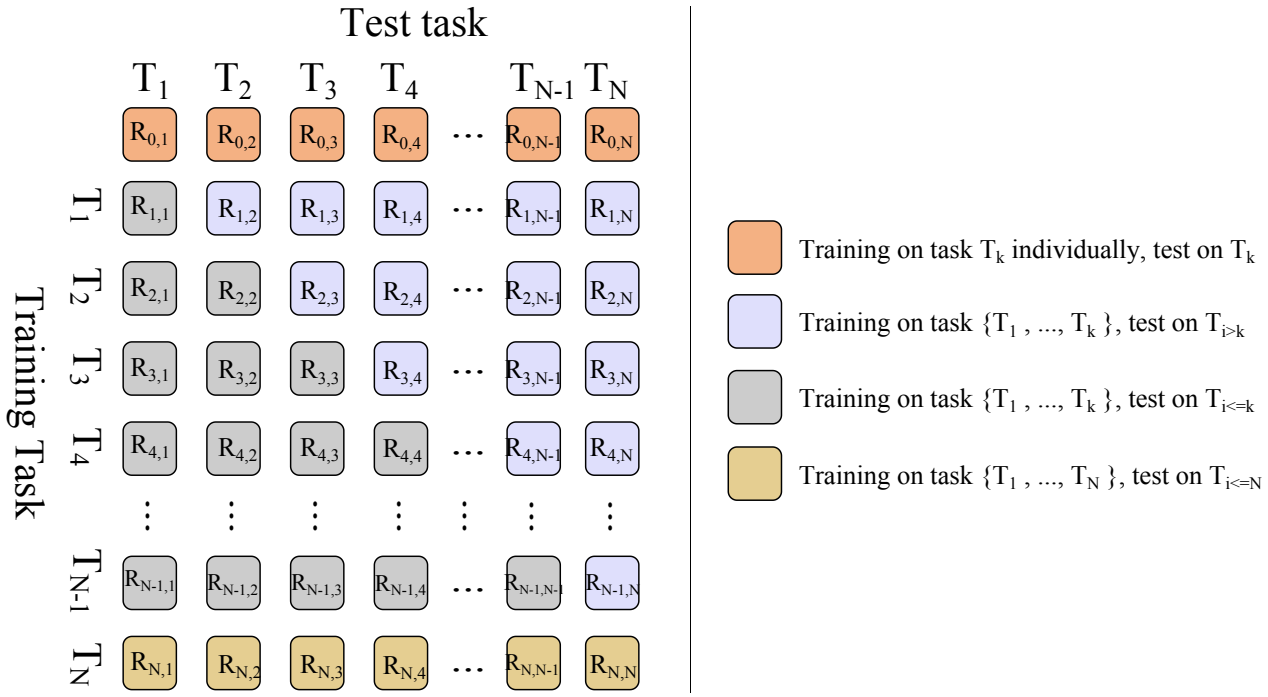

Fig. 18: Illustration of calculating metrics.

LCA at β is the area of the convergence curve Z_b as a function of $b \in [0, \beta]$:

$$LCA_\beta = \frac{1}{\beta+1}\int_0^\beta Z_b\, db = \frac{1}{\beta+1}\sum_{b=0}^{\beta} Z_b \tag{6}$$

The Learning Curve Area (LCA) provides insights into model learning dynamics. LCA_0 measures the average 0-shot performance, similar to forward transfer ([221]). LCA_β, quantifying the area under the Z_b curve, evaluates both average 0-shot performance and learning speed. Although two models may achieve similar Z_b or A_T values, they can differ significantly in LCA_β due to variations in learning rates. This metric is crucial for identifying models that quickly learn from few examples, particularly when β is small.

Qin et al. [23] propose two metrics designed to evaluate pre-trained language models (PLMs) based on their performance within learned domains: Average Perplexity (*AP*) and Average Increased Perplexity (AP^+). The aforementioned metrics are utilized to assess key capabilities of PLMs, such as instruction following and safety, as discussed in Wang et al. [121].

b. Memory Stability.

Memory stability is typically evaluated by backward transfer (BWT) [221] and forgetting measure (FM) [225].

Backward Transfer (BWT) emerges as a pivotal concept extensively documented in the literature, notably by Lopez et al. [221] and Wu et al. [226]. BWT measures the performance degradation on previously mastered tasks after the model is trained on new tasks. This performance degradation phenomenon, often referred to as "forgetting".

$$BWT = \frac{1}{N-1}\sum_{i=1}^{N-1} R_{N,i} - R_{i,i} \tag{7}$$

Chaudhry et al. [225] introduce the Forgetting Measure (FM), a metric designed to quantify the extent of forgetting a model experiences for a specific task. The forgetting for a given task T_j after sequential training on tasks up to T_N is quantified as the difference between the highest proficiency ($max(R_{l,j})$) achieved on task T_j during initial training and its proficiency ($R_{N,j}$) after subsequent learning phases:

$$f_j = \max_{l\in\{1,\ldots,N-1\}} (R_{l,j} - R_{N,j}), \quad \forall j < N. \tag{8}$$

For the purpose of quantifying forgetting in previous tasks, the function f_j is defined within the interval $[-1, 1]$ for $j < N$.

Furthermore, to account for the number of tasks previously encountered, the Forgetting Measure (FM) at the N-th task represents the mean level of forgetting across all preceding tasks:

$$FM = \frac{1}{N-1}\sum_{j=1}^{N-1} f_j \tag{9}$$

A lower FM indicates better retention of previous tasks. Here, the *expansion* or $R_{j,j}$ serves as a more effective quantifier of retained knowledge concerning past tasks, as opposed to using *max*. Nonetheless, *max* remains a valuable estimator for assessing the extent of forgetting that occurs throughout the learning process.

Davari et al. [227] propose a method named linear probes (LP) to assess representation forgetting. This approach measures the effectiveness of learned representations via an optimal linear classifier trained on the frozen activations of a base network. Representation forgetting is quantified by evaluating the change in Language Processing (LP) performance before and after the introduction of a new task. Formally, for each model (f_{θ_i}) at time step i of a task sequence, the classifier (W_i^*) is optimized as: $W_i^* = \arg\min_{W_i} \mathscr{L}(W_i; f_{\theta_i}(X_i), Y_i)$, where $\mathscr{L}$, X_i, and Y_i represent the objective function, input data, and labels for task i, respectively. The degree of representational forgetting between two model states, θ_a and θ_b,

where θ_b is derived later in the sequence, is evaluated by calculating the difference in scores: $Score(W_a f_{\theta_a}(X_a), Y_a) - Score(W_b f_{\theta_b}(X_a), Y_a)$, where *Score* represents the performance metric, such as accuracy, on the task.

Kemker et al. [228] introduce three metrics designed to CF: Ω_{base}, Ω_{new}, and Ω_{all}. Ω_{base} assesses retention of initial learning, Ω_{new} measures recall of new tasks, and Ω_{all} evaluates overall proficiency in maintaining old knowledge and acquiring new information.

$$\Omega_{\text{base}} = \frac{1}{N-1}\sum_{i=2}^{N}\frac{\alpha_{\text{base},i}}{\alpha_{\text{ideal}}} \tag{10}$$

$$\Omega_{\text{new}} = \frac{1}{N-1}\sum_{i=2}^{N}\frac{\alpha_{\text{new},i}}{\alpha_{\text{ideal}}} \tag{11}$$

$$\Omega_{\text{all}} = \frac{1}{N-1}\sum_{i=2}^{N}\frac{\alpha_{\text{all},i}}{\alpha_{\text{ideal}}} \tag{12}$$

where N represents the total number of sessions, $\alpha_{\text{new},i}$ is the test accuracy after learning session i, $\alpha_{\text{base},i}$ denotes the accuracy on the initial session after i sessions, and $\alpha_{\text{all},i}$ refers to the test accuracy across all test data for classes encountered up to point i. The ideal performance (α_{ideal}) is defined as the offline MLP accuracy on the base set. To facilitate comparative analysis across different datasets, Ω_{base} and Ω_{all} are normalized by α_{ideal}. Consequently, unless a model surpasses α_{ideal}, normalized results will range from 0 to 1, enabling consistent cross-dataset comparisons.

Additionally, researchers [139] devise a novel metric, termed the Knowledge Loss Ratio (KLR), quantifies knowledge degradation using principles from information theory.

c. Learning Plasticity.

Evaluating learning plasticity can be effectively accomplished through two key metrics: forward transfer (FWT) [221] and intransigence measure (IM) [225].

Forward Transfer (FWT) [221] assesses the beneficial effects on the performance of subsequent tasks following a model's training on prior tasks.

$$FWT = \frac{1}{N-1}\sum_{i=2}^{N} R_{i-1,i} - R_{0,i} \tag{13}$$

where $R_{0,i}$ denotes the performance metric associated with training on task i independently. Higher values of FWT indicate superior model performance. It is important to note that discussing backward transfer for the initial task is not applicable, as there are no preceding tasks to influence its performance.

Intransigence measure (IM), as defined by Chaudhry et al. [225], quantifies a model's inability to learn new tasks. This measure is calculated by comparing the performance difference of a task when trained jointly with other tasks versus when trained in a continual learning setting. Then the intransigence for the N-th task can be defined as:

$$IM = R_N^* - R_{N,N}, \tag{14}$$

Where R_N^* represents the accuracy achieved on the held-out dataset of the N-th task, $R_{N,N}$ indicates the accuracy on the N-th task upon completion of training in an incremental sequence up to and including task N. Note, $IM_N \in [-1,1]$, and lower values indicate superior performance.

Moreover, Koh et al. [139] introduce novel metrics, known as Knowledge Gain Ratio (KGR), which quantifies the capacity to acquire new knowledge through the calculation of knowledge gain.

d. Metrics for Continual Pre-training.

CKL [40] introduces a novel metric, named FUAR (FORGOTTEN / (UPDATED + ACQUIRED) RATIO), which quantitatively measures the efficiency of each CKL method. It calculates the number of instances of time-invariant knowledge that a model forgets in order to learn or update one instance of new knowledge. When FUAR is equal to 1.0, it signifies an equilibrium where one time-invariant knowledge instance is forgotten on average to obtain a new or updated knowledge instance. Formally, FUAR is defined as:

$$Eq_1 = \sum_{i=0}^{N-1}\max(0, \text{Gap}(T_i^F, D_i, D_N))\mathbb{1}_{\{T_i^F \neq n.d.\}} \tag{15}$$

$$\begin{aligned} Eq_2 = &\sum_{i=0}^{N-1}\max(0, \text{Gap}(T_B^U, D_N, D_i))\mathbb{1}_{\{T_i^F \neq n.d.\}} \\ &+ \max(0, \text{Gap}(T_N^A, D_N, D_i))\mathbb{1}_{\{T_i^F \neq n.d.\}} \end{aligned} \tag{16}$$

$$FUAR(\mathbb{T}^F, T_N^U, T_N^A) = \begin{cases} \dfrac{Eq_1}{Eq_2} & \text{if } denominator > 0 \\ no\ gain & \text{otherwise} \end{cases} \tag{17}$$

where T represents an arbitrary task, and $(D_i)_{i=0}^N$ is a sequence of corpora for LM pretraining. $\text{Gap}(T, D_a, D_b)$ is $\text{Score}(T)$ of LM_a - $\text{Score}(T)$ of LM_b, where LM_a is pretrained on D_a. $\mathbb{T}^F = (T_i^F)_{i=0}^{N-1}$ measures forgetting of invariant-knowledge from $(D_i)_{i=0}^{N-1}$. If no task is from D_i, T_i^F is "n.d." (not defined). T_N^U and T_N^A from D_N measure update and acquisition of new knowledge, respectively.

e. Online CL-Specific Metrics.

Near-future accuracy (NFA) [229] is introduced as a novel evaluation metric for OCL problem. Unlike traditional evaluation methods that assess models on immediately subsequent samples, NFA evaluates models on samples slightly further into the future, using a minimal shift S. Such operation can mitigate label correlation effects, which can adversely impact the accuracy of model adaptability assessments. The smallest shift S is selected to ensure that the test sample aligns closely with the distribution of recently observed training data. The calculation of NFA involves first checking if the model correctly predicts the label of a future sample, which can be expressed as $a_t = \mathbb{1}\{f_{\theta_t}(x_{t+1+S}) = y_{t+1+S}\}$. Subsequently, the running average is updated using the formula $A^{RA}t = \frac{1}{t}(A^{RA}t - 1 \cdot (t-1) + a_t)$.

Yogatama et al. [230] proposed a novel online codelength ($\ell(D)$), inspired by prequential encoding [231], to quantify how quickly an existing model can adapt to a new task.

$$\ell(D) = \log_2|y| - \sum_{i=2}^{N}\log_2 p(y_i|x_i;\theta_{D_{i-1}}) \tag{18}$$

where $|Y|$ is the number of possible labels (classes), and θ_{D_i} represents a particular subset of the dataset D. Similar to the approach in Latent Contextual Allocation (LCA) [224], the concept of *online codelength* is associated with the area under the learning curve.

VIII. Challenges and Further Work

Autonomous Continual Learning. Most existing studies in the domain of continual learning assume static datasets with known distributions in a relatively closed environment. Moreover, these studies mainly focus on simple tasks (e.g., text classification, sentiment analysis and intent classification) with clear labels. These assumptions do not hold in real-world applications, where environments continually evolve and introduce novel stimuli. A key challenge is to develop continual learning models that operate effectively in complex, noisy environments where clear labels are not always available, and task domains frequently change. Liu et al. [232] recently proposed the SOLA framework to address these limitations by facilitating autonomous adaptation in AI systems. Despite this progress, significant challenges remain in enabling these systems to independently adjust to new, dynamic environments without ongoing human oversight. Future research should focus on developing algorithms capable of autonomously detecting and adapting to shifts in data distribution, thereby improving the applicability of AI in dynamic real-world scenarios.

Learning Knowledge from Conversation. Traditional AI systems are typically trained on static data sets, which starkly contrasts with human conversational learning that dynamically updates knowledge through interaction [50]. The challenge for AI lies in transitioning from static data learning to more dynamic, conversational engagements. The future direction in this area could involve the development of models that mimic human conversational learning processes, capable of context adaptation, new concept inference, and dynamic knowledge application within ongoing interactions.

Multi-modal Continual Learning. Continual learning research has predominantly concentrated on natural language processing tasks such as sentiment analysis and text classification. Recent studies have begun exploring basic multimodal tasks, such as text-to-image retrieval, text-image classification, and visual question answering. The integration of diverse data types—textual, visual, and auditory—poses a substantial challenge. Future studies should expand to more complex multi-modal datasets and strive to devise methodologies that effectively synthesize these varied modalities, thereby enhancing the model's capability to maintain continuous learning across different sensory inputs.

Privacy Protection in Continual Learning. Privacy protection in continual learning systems poses a significant challenge, particularly as these systems are designed to continuously update and refine their models based on incoming data streams. Unlike traditional static machine learning models, continual learning systems frequently access and process sensitive data across different contexts and time periods, raising substantial concerns about data confidentiality and user privacy. Effective privacy-preserving mechanisms must be integrated into the architecture of these systems to ensure that they do not inadvertently expose or misuse personal data. Techniques such as differential privacy [233], federated learning [234], and secure multi-party computation [235] offer promising solutions by allowing models to learn from decentralized data sources without needing to access the actual data directly. Future research in continual learning should not only focus on enhancing learning efficiency and adaptability but also prioritize the development of robust frameworks that safeguard user privacy across all phases of data handling and model updating.

Robust Continual Learning. The existing studies mainly focus on designing a continual learning model to improve the performance of forgetting and transferring with various metrics while the robustness of continual learning systems is not well studied. It is critical, especially in applications where safety and reliability are paramount. The main challenges include evaluating the robustness of these systems against adversarial attacks or when faced with drastically changing environments. Future research could focus on developing evaluation metrics for robustness in continual learning and designing systems that maintain performance reliability over time despite environmental changes.

Large-Scale and High-Quality Datasets and Benchmarks. As discussed in Section VI, most of the datasets are constructed by merging the existing datasets. This often results in datasets that lack diversity and real-world complexity, which hampers the development of robust and adaptable continual learning models. The creation of large-scale, high-quality datasets that accurately reflect real-world complexities represents a critical challenge. Moving forward, the development of such datasets and benchmarks will be essential not only for assessing the efficacy of continual learning algorithms but also for pushing the limits of what these algorithms can achieve in practical settings.

IX. Conclusions

This survey provides an in-depth exploration of continual learning (CL) methodologies tailored for foundation language models (LMs), such as pre-trained language models (PLMs), large language models (LLMs), and vision-language models (VLMs). By integrating the dynamic adaptability of CL with the robust foundational capabilities of LMs, this field promises to significantly advance the state of artificial intelligence. We categorize existing research into offline and online continual learning paradigms, offering a clear distinction between the settings and methodologies used within these frameworks. Offline CL is discussed in terms of domain-incremental, task-incremental, and class-incremental learning. Meanwhile, online CL is analyzed with a focus on the delineation between hard and blurry task boundaries, providing insights into how these approaches handle real-time data streams. Our review of the literature not only clarifies the current landscape of CL approaches for foundation LMs but also emphasizes the innovative integration of continual pre-training, parameter-efficient tuning, and instruction tuning methods that are specifically designed to

leverage the vast capabilities of foundation LMs. Furthermore, we highlight the main characteristics of datasets used in this domain and the metrics that effectively measure both the mitigation of catastrophic forgetting and the enhancement of knowledge transfer. This work hopes to inspire further research that will ultimately lead to more robust, efficient, and intelligent systems capable of lifelong learning.

REFERENCES

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[3] J. Zhou, P. Ke, X. Qiu, M. Huang, and J. Zhang, "Chatgpt: potential, prospects, and limitations," *Frontiers of Information Technology & Electronic Engineering*, pp. 1–6, 2023.

[4] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, vol. 20, no. 4, pp. 447–482, 2023.

[5] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 5436–5443, survey Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/762

[6] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[9] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[14] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[15] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[16] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[17] G. M. Van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1185–1197, 2022.

[18] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks*, vol. 113, pp. 54–71, 2019.

[19] D.-W. Zhou, Q.-W. Wang, Z.-H. Qi, H.-J. Ye, D.-C. Zhan, and Z. Liu, "Deep class-incremental learning: A survey," *arXiv preprint arXiv:2302.03648*, 2023.

[20] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[21] C. Qin and S. Joty, "LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=HCRVf71PMF

[22] Z. Ke and H. Xu, "Adapting bert for continual learning of a sequence of aspect sentiment classification tasks," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[23] Y. Qin, J. Zhang, Y. Lin, Z. Liu, P. Li, M. Sun, and J. Zhou, "Elle: Efficient lifelong pre-training for emerging data," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2789–2810.

[24] A. Madotto, Z. Lin, Z. Zhou, S. Moon, P. Crook, B. Liu, Z. Yu, E. Cho, and Z. Wang, "Continual learning in task-oriented dialogue systems," *arXiv preprint arXiv:2012.15504*, 2020.

[25] D. Li, Z. Chen, E. Cho, J. Hao, X. Liu, F. Xing, C. Guo, and Y. Liu, "Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5441–5454.

[26] B. Geng, F. Yuan, Q. Xu, Y. Shen, R. Xu, and M. Yang, "Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 517–523. [Online]. Available: https://aclanthology.org/2021.acl-short.66

[27] Z. Ke, B. Liu, H. Xu, and L. Shu, "Classic: Continual and contrastive learning of aspect sentiment classification tasks," in *Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, 2021.

[28] Z. Ke, H. Lin, Y. Shao, H. Xu, L. Shu, and B. Liu, "Continual training of language models for few-shot learning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 10 205–10 216.

[29] Q. Zhu, B. Li, F. Mi, X. Zhu, and M. Huang, "Continual prompt tuning for dialog state tracking," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1124–1137. [Online]. Available: https://aclanthology.org/2022.acl-long.80

[30] G. Castellucci, S. Filice, D. Croce, and R. Basili, "Learning to solve nlp tasks in an incremental number of languages," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 837–847. [Online]. Available: https://aclanthology.org/2021.acl-short.106

[31] X. Cheng, Y. Lin, X. Chen, D. Zhao, and R. Yan, "Decouple knowledge from paramters for plug-and-play language modeling," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 288–14 308. [Online]. Available: https://aclanthology.org/2023.findings-acl.901

[32] A. Cossu, T. Tuytelaars, A. Carta, L. Passaro, V. Lomonaco, and D. Bacciu, "Continual pre-training mitigates forgetting in language and vision," *arXiv preprint arXiv:2205.09357*, 2022.

[33] S. Lee, "Toward continual learning for conversational agents," *arXiv preprint arXiv:1712.09943*, 2017.

[34] Z. Ke, Y. Shao, H. Lin, T. Konishi, G. Kim, and B. Liu, "Continual pre-training of language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=m_GDIItaI3o

[35] H. Zhang, L. Gui, Y. Zhai, H. Wang, Y. Lei, and R. Xu, "Copf: Continual learning human preference through optimal policy fitting," *arXiv preprint arXiv:2310.15694*, 2023.

[36] F. Sun, C. Ho, and H. Lee, "LAMOL: language modeling for lifelong language learning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=Skgxcn4YDS

[37] H. Wang, R. Fu, X. Zhang, and J. Zhou, "Rvae-lamol: Residual variational autoencoder to enhance lifelong language learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–9.

[38] Y. Zhang, Y. Wang, F. Cheng, S. Kurohashi *et al.*, "Reformulating domain adaptation of large language models as adapt-retrieve-revise," *arXiv preprint arXiv:2310.03328*, 2023.

[39] W. Chen, Y. Zhou, N. Du, Y. Huang, J. Laudon, Z. Chen, and C. Cui, "Lifelong language pretraining with distribution-specialized experts," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5383–5395.

[40] J. Jang, S. Ye, S. Yang, J. Shin, J. Han, G. Kim, S. J. Choi, and M. Seo, "Towards continual knowledge learning of language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=vfsRB5MImo9

[41] Y. Xie, K. Aggarwal, and A. Ahmad, "Efficient continual pre-training for building domain specific large language models," *arXiv preprint arXiv:2311.08545*, 2023.

[42] H. Zhang, Y. Lei, L. Gui, M. Yang, Y. He, H. Wang, and R. Xu, "Cppo: Continual learning for reinforcement learning with human feedback," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=86zAUE80pP

[43] S. Ma, S. Huang, S. Huang, X. Wang, Y. Li, H.-T. Zheng, P. Xie, F. Huang, and Y. Jiang, "Ecomgpt-ct: Continual pre-training of e-commerce large language models with semi-structured data," *arXiv preprint arXiv:2312.15696*, 2023.

[44] Y. Wang, Z. Huang, and X. Hong, "S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5682–5695, 2022.

[45] H. Yi, Z. Qin, Q. Lao, W. Xu, Z. Jiang, D. Wang, S. Zhang, and K. Li, "Towards general purpose medical ai: Continual learning medical foundation model," *arXiv preprint arXiv:2303.06580*, 2023.

[46] X. Zhang, F. Zhang, and C. Xu, "Vqacl: A novel visual question answering continual learning setting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 102–19 112.

[47] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, and A. Almahairi, "Progressive prompts: Continual learning for language models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=UJTgQBc91_

[48] Z. Ke, B. Liu, N. Ma, H. Xu, and L. Shu, "Achieving forgetting prevention and knowledge transfer in continual learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 22 443–22 456. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/bcd0049c35799cdf57d06eaf2eb3cff6-Paper.pdf

[49] C. Wang, H. Pan, Y. Liu, K. Chen, M. Qiu, W. Zhou, J. Huang, H. Chen, W. Lin, and D. Cai, "Mell: Large-scale extensible user intent classification for dialogue systems with meta lifelong learning," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3649–3659.

[50] B. Liu and S. Mazumder, "Lifelong and continual learning dialogue systems: learning during conversation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 15 058–15 063.

[51] C. Qin and S. Joty, "Continual few-shot relation learning via embedding space regularization and data augmentation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2776–2789.

[52] Y. Zhao, Y. Zheng, Z. Tian, C. Gao, J. Sun, and N. L. Zhang, "Prompt conditioned vae: Enhancing generative replay for lifelong learning in task-oriented dialogue," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 153–11 169. [Online]. Available: https://aclanthology.org/2022.emnlp-main.766

[53] X. Jin, B. Y. Lin, M. Rostami, and X. Ren, "Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 714–729. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.62

[54] W. Yin, J. Li, and C. Xiong, "ConTinTin: Continual learning from task instructions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3062–3072. [Online]. Available: https://aclanthology.org/2022.acl-long.218

[55] A. Maekawa, H. Kamigaito, K. Funakoshi, and M. Okumura, "Generative replay inspired by hippocampal memory indexing for continual language learning," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 930–942.

[56] Y. Zhang, X. Wang, and D. Yang, "Continual sequence generation with adaptive compositional modules," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 3653–3667. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.255

[57] J. Mok, J. Do, S. Lee, T. Taghavi, S. Yu, and S. Yoon, "Large-scale lifelong learning of in-context instructions and how to tackle it," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 573–12 589. [Online]. Available: https://aclanthology.org/2023.acl-long.703

[58] F. Yuan, G. Zhang, A. Karatzoglou, J. Jose, B. Kong, and Y. Li, "One person, one model, one world: Learning continual user representation without forgetting," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 696–705. [Online]. Available: https://doi.org/10.1145/3404835.3462884

[59] S. Kim, N. Lee, D. Kim, M. Yang, and C. Park, "Task relation-aware continual user representation learning," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1107–1119. [Online]. Available: https://doi.org/10.1145/3580305.3599516

[60] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8968–8975.

[61] Y. Qin, C. Qian, X. Han, Y. Lin, H. Wang, R. Xie, Z. Liu, M. Sun, and J. Zhou, "Recyclable tuning for continual pre-training," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 11 403–11 426.

[62] C. Song, X. Han, Z. Zeng, K. Li, C. Chen, Z. Liu, M. Sun, and T. Yang, "Conpet: Continual parameter-efficient tuning for large language models," *arXiv preprint arXiv:2309.14763*, 2023.

[63] S. Cahyawijaya, H. Lovenia, T. Yu, W. Chung, and P. Fung, "InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning," in *Proceedings of the First Workshop in South East Asian Language Processing*, D. Wijaya, A. F. Aji, C. Vania, G. I. Winata, and A. Purwarianti, Eds. Nusa Dua, Bali, Indonesia: Association for Computational Linguistics, Nov. 2023, pp. 55–78. [Online]. Available: https://aclanthology.org/2023.sealp-1.5

[64] T. Scialom, T. Chakrabarty, and S. Muresan, "Fine-tuned language models are continual learners," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6107–6122. [Online]. Available: https://aclanthology.org/2022.emnlp-main.410

[65] M. Du, A. T. Luu, B. Ji, and S.-k. Ng, "From static to dynamic: A continual learning framework for large language models," *arXiv preprint arXiv:2310.14248*, 2023.

[66] J. Jang, S. Kim, S. Ye, D. Kim, L. Logeswaran, M. Lee, K. Lee, and M. Seo, "Exploring the benefits of training expert language models over instruction tuning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 14 702–14 729.

[67] X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang, "Orthogonal subspace learning for language model continual learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 658–10 671. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.715

[68] B. PENG, Z. Tian, S. Liu, M.-C. Yang, and J. Jia, "Scalable language model with generalized continual learning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=mz8owj4DXu

[69] H. Zhu, Y. Wei, X. Liang, C. Zhang, and Y. Zhao, "Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 257–22 267.

[70] Z. Zheng, M. Ma, K. Wang, Z. Qin, X. Yue, and Y. You, "Preventing zero-shot transfer degradation in continual learning of vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 125–19 136.

[71] J. Yu, Y. Zhuge, L. Zhang, P. Hu, D. Wang, H. Lu, and Y. He, "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[72] Z. Qian, X. Wang, X. Duan, P. Qin, Y. Li, and W. Zhu, "Decouple before interact: Multi-modal prompt learning for continual visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2953–2962.

[73] Z. Wang, Y. Liu, T. Ji, X. Wang, Y. Wu, C. Jiang, Y. Chao, Z. Han, L. Wang, X. Shao *et al.*, "Rehearsal-free continual language learning via efficient parameter isolation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 10 933–10 946.

[74] Y. Huang, Y. Zhang, J. Chen, X. Wang, and D. Yang, "Continual learning for text classification with information disentanglement based regularization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 2736–2746. [Online]. Available: https://aclanthology.org/2021.naacl-main.218

[75] V. Varshney, M. Patidar, R. Kumar, L. Vig, and G. Shroff, "Prompt augmented generative replay via supervised contrastive learning for lifelong intent detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 1113–1127.

[76] C. Xia, W. Yin, Y. Feng, and P. Yu, "Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system," *arXiv preprint arXiv:2104.11882*, 2021.

[77] N. Monaikul, G. Castellucci, S. Filice, and O. Rokhlenko, "Continual learning for named entity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 570–13 577.

[78] G. Li, Y. Zhai, Q. Chen, X. Gao, J. Zhang, and Y. Zhang, "Continual few-shot intent detection," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 333–343.

[79] M. Wójcik, W. Kościukiewicz, M. Baran, T. Kajdanowicz, and A. Gonczarek, "Domain-agnostic neural architecture for class incremental continual learning in document processing platform," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, S. Sitaram, B. Beigman Klebanov, and J. D. Williams, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 527–537. [Online]. Available: https://aclanthology.org/2023.acl-industry.51

[80] T. Liu, L. Ungar, and J. Sedoc, "Continual learning for sentence representations using conceptors," 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3274–3279. [Online]. Available: https://aclanthology.org/N19-1331

[81] X. Cao, H. Lu, L. Huang, X. Liu, and M.-M. Cheng, "Generative multi-modal models are good class-incremental learners," *arXiv preprint arXiv:2403.18383*, 2024.

[82] J. Kim, Y. Ku, J. Kim, J. Cha, and S. Baek, "Vlm-pl: Advanced pseudo labeling approach class incremental object detection with vision-language model," *arXiv preprint arXiv:2403.05346*, 2024.

[83] X. Liu, X. Cao, H. Lu, J.-w. Xiao, A. D. Bagdanov, and M.-M. Cheng, "Class incremental learning with pre-trained vision-language models," *arXiv preprint arXiv:2310.20348*, 2023.

[84] D.-W. Zhou, Y. Zhang, J. Ning, H.-J. Ye, D.-C. Zhan, and Z. Liu, "Learning without forgetting for vision-language models," *arXiv preprint arXiv:2305.19270*, 2023.

[85] M. G. Z. A. Khan, M. F. Naeem, L. Van Gool, D. Stricker, F. Tombari, and M. Z. Afzal, "Introducing language guidance in prompt-based continual learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 463–11 473.

[86] S. Jha, D. Gong, and L. Yao, "Clap4clip: Continual learning with probabilistic finetuning for vision-language models," *arXiv preprint arXiv:2403.19137*, 2024.

[87] C. de Masson D'Autume, S. Ruder, L. Kong, and D. Yogatama, "Episodic memory in lifelong language learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[88] Z. Wang, S. V. Mehta, B. Poczos, and J. Carbonell, "Efficient meta lifelong-learning with limited memory," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 535–548. [Online]. Available: https://aclanthology.org/2020.emnlp-main.39

[89] N. Holla, P. Mishra, H. Yannakoudakis, and E. Shutova, "Meta-learning with sparse experience replay for lifelong language learning," *arXiv preprint arXiv:2009.04891*, 2020.

[90] Q. Liu, X. Yu, S. He, K. Liu, and J. Zhao, "Lifelong intent detection via multi-strategy rebalancing," *arXiv preprint arXiv:2108.04445*, 2021.

[91] Y. Shen, X. Zeng, and H. Jin, "A progressive model to enable continual learning for semantic slot filling," 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)*, 2019, pp. 1279–1284.

[92] Q. Wang, R. Wang, Y. Wu, X. Jia, and D. Meng, "Cba: Improving online continual learning via continual bias adaptor," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 082–19 092.

[93] J.-Y. Moon, K.-H. Park, J. U. Kim, and G.-M. Park, "Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 731–11 741.

[94] Z. Cui, Y. Peng, X. Wang, M. Zhu, and J. Zhou, "Continual vision-language retrieval via dynamic knowledge rectification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 11 704–11 712.

[95] S. V. Mehta, D. Patil, S. Chandar, and E. Strubell, "An empirical investigation of the role of pre-training in lifelong learning," 2021.

[96] K.-Y. Lee, Y. Zhong, and Y.-X. Wang, "Do pre-trained models benefit equally in continual learning?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6485–6493.

[97] T. Scialom, T. Chakrabarty, and S. Muresan, "Fine-tuned language models are continual learners," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 6107–6122.

[98] Z. Han, C. Gao, J. Liu, S. Q. Zhang *et al.*, "Parameter-efficient fine-tuning for large models: A comprehensive survey," *arXiv preprint arXiv:2403.14608*, 2024.

[99] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang, "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)*, 2022, pp. 61–68.

[100] N. Mundra, S. Doddapaneni, R. Dabre, A. Kunchukuttan, R. Puduppully, and M. M. Khapra, "A comprehensive analysis of adapter efficiency," in *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, 2024, pp. 136–154.

[101] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[102] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[103] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. Van De Weijer, "Class-incremental learning: survey and performance evaluation on image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5513–5533, 2022.

[104] E. Belouadah, A. Popescu, and I. Kanellos, "A comprehensive study of class incremental learning algorithms for visual tasks," *Neural Networks*, vol. 135, pp. 38–54, 2021.

[105] T. L. Hayes, G. P. Krishnan, M. Bazhenov, H. T. Siegelmann, T. J. Sejnowski, and C. Kanan, "Replay in deep learning: Current approaches and missing biological elements," *Neural computation*, vol. 33, no. 11, pp. 2908–2950, 2021.

[106] H. Qu, H. Rahmani, L. Xu, B. Williams, and J. Liu, "Recent advances of continual learning in computer vision: An overview," *arXiv preprint arXiv:2109.11369*, 2021.

[107] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.

[108] M. Biesialska, K. Biesialska, and M. R. Costa-jussà, "Continual lifelong learning in natural language processing: A survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6523–6541.

[109] Z. Ke and B. Liu, "Continual learning of natural language processing tasks: A survey," *arXiv preprint arXiv:2211.12701*, 2022.

[110] P. Zhang and S. Kim, "A survey on incremental update for neural recommender systems," *arXiv preprint arXiv:2303.02851*, 2023.

[111] K. Shaheen, M. A. Hanif, O. Hasan, and M. Shafique, "Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks," *Journal of Intelligent & Robotic Systems*, vol. 105, no. 1, p. 9, 2022.

[112] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information fusion*, vol. 58, pp. 52–68, 2020.

[113] S. Gururangan, M. Lewis, A. Holtzman, N. A. Smith, and L. Zettlemoyer, "Demix layers: Disentangling domains for modular language modeling," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5557–5576.

[114] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[115] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, and J. Pfeiffer, "Adapters: A unified library for parameter-efficient and modular transfer learning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 149–160.

[116] E. Gogoulou, T. Lesort, M. Boman, and J. Nivre, "A study of continual learning under language shift," *arXiv preprint arXiv:2311.01200*, 2023.

[117] D. Cheng, S. Huang, and F. Wei, "Adapting large language models via reading comprehension," *arXiv preprint arXiv:2309.09530*, 2023.

[118] H. Wang, W. Xiong, M. Yu, X. Guo, S. Chang, and W. Y. Wang, "Sentence embedding alignment for lifelong relation extraction," 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 796–806. [Online]. Available: https://aclanthology.org/N19-1086

[119] S. Mazumder and B. Liu, *Lifelong and Continual Learning Dialogue Systems*. Springer Nature, 2024.

[120] B. Liu, "Learning on the job: Online lifelong and continual learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 09, 2020, pp. 13 544–13 549.

[121] X. Wang, Y. Zhang, T. Chen, S. Gao, S. Jin, X. Yang, Z. Xi, R. Zheng, Y. Zou, T. Gui, Q. Zhang, and X. Huang, "Trace: A comprehensive benchmark for continual learning in large language models," 2024. [Online]. Available: https://openreview.net/forum?id=xelrLobW0n

[122] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning," *arXiv e-prints*, p. arXiv:2308.08747, Aug. 2023.

[123] C. Chen, J. Zhu, X. Luo, H. Shen, L. Gao, and J. Song, "Coin: A benchmark of continual instruction tuning for multimodel large language model," *arXiv preprint arXiv:2403.08350*, 2024.

[124] U. Michieli, P. P. Parada, and M. Ozay, "Online continual learning in keyword spotting for low-resource devices via pooling high-order temporal statistics," *arXiv preprint arXiv:2307.12660*, 2023.

[125] S. Vander Eeckt *et al.*, "Rehearsal-free online continual learning for automatic speech recognition," *arXiv e-prints*, pp. arXiv–2306, 2023.

[126] K. Javed and M. White, "Meta-learning representations for continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[127] S. Beaulieu, L. Frati, T. Miconi, J. Lehman, K. O. Stanley, J. Clune, and N. Cheney, "Learning to continually learn," *arXiv preprint arXiv:2002.09571*, 2020.

[128] H. Yin, P. Li *et al.*, "Mitigating forgetting in online continual learning with neuron calibration," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 260–10 272, 2021.

[129] Y. Gu, X. Yang, K. Wei, and C. Deng, "Not just selection, but exploration: Online class-incremental continual learning via dual view consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7442–7451.

[130] Y. Guo, B. Liu, and D. Zhao, "Online continual learning through mutual information maximization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8109–8126.

[131] J. Bang, H. Koh, S. Park, H. Song, J.-W. Ha, and J. Choi, "Online continual learning on a contaminated data stream with blurry task boundaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9275–9284.

[132] R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia, "Online continual learning with maximal interfered retrieval," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/15825aee15eb335cc13f9b559f166ee8-Paper.pdf

[133] D. Shim, Z. Mai, J. Jeong, S. Sanner, H. Kim, and J. Jang, "Online class-incremental continual learning with adversarial shapley value," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9630–9638.

[134] L. Caccia, R. Aljundi, N. Asadi, T. Tuytelaars, J. Pineau, and E. Belilovsky, "New insights on reducing abrupt representation change in online continual learning," *arXiv preprint arXiv:2104.05025*, 2021.

[135] H. Lin, B. Zhang, S. Feng, X. Li, and Y. Ye, "Pcr: Proxy-based contrastive replay for online class-incremental continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 246–24 255.

[136] Y. Zhang, B. Pfahringer, E. Frank, A. Bifet, N. J. S. Lim, and Y. Jia, "A simple but strong baseline for online continual learning: Repeated augmented rehearsal," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 771–14 783, 2022.

[137] J. He, R. Mao, Z. Shao, and F. Zhu, "Incremental learning in online scenario," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 926–13 935.

[138] E. Fini, S. Lathuiliere, E. Sangineto, M. Nabi, and E. Ricci, "Online continual learning under extreme memory constraints," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 720–735.

[139] H. Koh, M. Seo, J. Bang, H. Song, D. Hong, S. Park, J.-W. Ha, and J. Choi, "Online boundary-free continual learning by scheduled data prior," in *The Eleventh International Conference on Learning Representations*, 2022.

[140] X. Jin, A. Sadhu, J. Du, and X. Ren, "Gradient-based editing of memory examples for online task-free continual learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 193–29 205, 2021.

[141] Y. Guo, B. Liu, and D. Zhao, "Dealing with cross-task class discrimination in online continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 878–11 887.

[142] H.-J. Chen, A.-C. Cheng, D.-C. Juan, W. Wei, and M. Sun, "Mitigating forgetting in online continual learning via instance-aware parameterization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 466–17 477, 2020.

[143] J. Yoon, D. Madaan, E. Yang, and S. J. Hwang, "Online coreset selection for rehearsal-based continual learning," *arXiv preprint arXiv:2106.01085*, 2021.

[144] A. Ororbia, A. Mali, C. L. Giles, and D. Kifer, "Lifelong neural predictive coding: Learning cumulatively online without forgetting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5867–5881, 2022.

[145] H. Koh, D. Kim, J.-W. Ha, and J. Choi, "Online continual learning on class incremental blurry task configuration with anytime inference," *arXiv preprint arXiv:2110.10031*, 2021.

[146] N. Gunasekara, B. Pfahringer, H. M. Gomes, and A. Bifet, "Survey on online streaming continual learning," in *Proceedings of IJCAI*, 2023.

[147] Y. Wei, J. Ye, Z. Huang, J. Zhang, and H. Shan, "Online prototype learning for online continual learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 764–18 774.

[148] A. Chrysakis and M.-F. Moens, "Online bias correction for task-free continual learning," *ICLR 2023 at OpenReview*, 2023.

[149] R. Semola, V. Lomonaco, and D. Bacciu, "Continual-learning-as-a-service (claas): On-demand efficient adaptation of predictive models," *arXiv preprint arXiv:2206.06957*, 2022.

[150] P. Yadav, Q. Sun, H. Ding, X. Li, D. Zhang, M. Tan, X. Ma, P. Bhatia, R. Nallapati, M. K. Ramanathan *et al.*, "Exploring continual learning for code generation models," *arXiv preprint arXiv:2307.02435*, 2023.

[151] Z. Zhang, M. Fang, L. Chen, and M.-R. Namazi-Rad, "Citb: A benchmark for continual instruction tuning," *arXiv preprint arXiv:2310.14510*, 2023.

[152] X. Jin, D. Zhang, H. Zhu, W. Xiao, S.-W. Li, X. Wei, A. Arnold, and X. Ren, "Lifelong pretraining: Continually adapting language models to emerging corpora," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 4764–4780.

[153] H. Hu, O. Sener, F. Sha, and V. Koltun, "Drinking from a firehose: Continual learning with web-scale natural language," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5684–5696, 2022.

[154] J. Jang, S. Ye, C. Lee, S. Yang, J. Shin, J. Han, G. Kim, and M. Seo, "Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 6237–6250.

[155] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[156] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 353–355.

[157] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[158] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: https://aclanthology.org/P11-1015

[159] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," in *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2017, pp. 364–371.

[160] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine learning proceedings 1995*. Elsevier, 1995, pp. 331–339.

[161] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang *et al.*, "An evaluation dataset for intent classification and out-of-scope prediction," *arXiv preprint arXiv:1909.02027*, 2019.

[162] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, "Benchmarking natural language understanding services for building conversational agents," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Springer, 2021, pp. 165–183.

[163] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders," *arXiv preprint arXiv:2003.04807*, 2020.

[164] J. Xu, B. Xu, P. Wang, S. Zheng, G. Tian, and J. Zhao, "Self-taught convolutional neural networks for short text clustering," *Neural Networks*, vol. 88, pp. 22–31, 2017.

[165] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan, "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8689–8696.

[166] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," *arXiv preprint arXiv:1810.00278*, 2018.

[167] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, "Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," *arXiv preprint arXiv:1810.10147*, 2018.

[168] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[169] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," *arXiv preprint arXiv:1810.07942*, 2018.

[170] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," *arXiv preprint arXiv:1810.13327*, 2018.

[171] F.-L. Li, M. Qiu, H. Chen, X. Wang, X. Gao, J. Huang, J. Ren, Z. Zhao, W. Zhao, L. Wang *et al.*, "Alime assist: An intelligent assistant for creating an innovative e-commerce experience," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2495–2498.

[172] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[173] Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "Automated rule selection for aspect extraction in opinion mining," in *Twenty-Fourth international joint conference on artificial intelligence*, 2015.

[174] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 231–240.

[175] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-snli: Natural language inference with natural language explanations," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[176] R. Bin Tareaf, "Tweets Dataset - Top 20 most followed users in Twitter social platform," 2017. [Online]. Available: https://doi.org/10.7910/DVN/JBXKFD

[177] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[178] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering. arxiv 2018," *arXiv preprint arXiv:1806.08730*.

[179] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[180] M. Volske, M. Potthast, S. Syed, and B. Stein, "Tl;dr: Mining reddit to learn automatic summarization," in *Proceedings of the Workshop on New Frontiers in Summarization*, L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 59–63. [Online]. Available: https://aclanthology.org/W17-4508

[181] K. Ethayarajh, Y. Choi, and S. Swayamdipta, "Understanding dataset difficulty with v-usable information," in *International Conference on Machine Learning*. PMLR, 2022, pp. 5988–6008.

[182] J. Novikova, O. Dušek, and V. Rieser, "The e2e dataset: New challenges for end-to-end generation," *arXiv preprint arXiv:1706.09254*, 2017.

[183] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *arXiv preprint arXiv:1508.01745*, 2015.

[184] V. Zhong, C. Xiong, and R. Socher, "Seq2sql: Generating structured queries from natural language using reinforcement learning," *arXiv preprint arXiv:1709.00103*, 2017.

[185] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[186] S. Iyer, I. Konstas, A. Cheung, and L. Zettlemoyer, "Mapping language to code in programmatic context," *arXiv preprint arXiv:1808.09588*, 2018.

[187] H. Husain, H.-H. Wu, T. Gazit, M. Allamanis, and M. Brockschmidt, "Codesearchnet challenge: Evaluating the state of semantic code search," *arXiv preprint arXiv:1909.09436*, 2019.

[188] S. Lu, D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. Clement, D. Drain, D. Jiang, D. Tang *et al.*, "Codexglue: A machine learning benchmark dataset for code understanding and generation," *arXiv preprint arXiv:2102.04664*, 2021.

[189] M. Tufano, C. Watson, G. Bavota, M. D. Penta, M. White, and D. Poshyvanyk, "An empirical study on learning bug-fixing patches in the wild via neural machine translation," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 28, no. 4, pp. 1–29, 2019.

[190] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[191] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: The 90% solution," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, R. C. Moore, J. Bilmes, J. Chu-Carroll, and M. Sanderson, Eds. New York City, USA: Association for Computational Linguistics, Jun. 2006, pp. 57–60. [Online]. Available: https://aclanthology.org/N06-2015

[192] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," *arXiv preprint arXiv:1506.02075*, 2015.

[193] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," *arXiv preprint arXiv:1904.02232*, 2019.

[194] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, "S2orc: The semantic scholar open research corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4969–4983. [Online]. Available: https://aclanthology.org/2020.acl-main.447

[195] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[196] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," *Advances in neural information processing systems*, vol. 32, 2019.

[197] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.

[198] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, B. Goodrich, A. Dubey, A. Cedilnik, and K.-Y. Kim, "Taskmaster-1: Toward a realistic and diverse dialog dataset," *arXiv preprint arXiv:1909.05358*, 2019.

[199] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap *et al.*, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks," *arXiv preprint arXiv:2204.07705*, 2022.

[200] F. Yuan, X. He, A. Karatzoglou, and L. Zhang, "Parameter-efficient transfer from sequential behaviors for user modeling and recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 1469–1478.

[201] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[202] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. [Online]. Available: https://aclanthology.org/P17-1147

[203] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, "Quac: Question answering in context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2174–2184. [Online]. Available: https://aclanthology.org/D18-1241

[204] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1965–1973.

[205] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[206] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[207] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[208] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International workshop ontoImage*, vol. 2, 2006.

[209] A. Yang, J. Lin, R. Men, C. Zhou, L. Jiang, X. Jia, A. Wang, J. Zhang, J. Wang, Y. Li *et al.*, "M6-t: Exploring sparse expert models and beyond," *arXiv preprint arXiv:2105.15082*, 2021.

[210] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[211] Z. Ni, L. Wei, S. Tang, Y. Zhuang, and Q. Tian, "Continual vision-language representation learning with off-diagonal information," in *International Conference on Machine Learning*. PMLR, 2023, pp. 26 129–26 149.

[212] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, "Cl-masr: A continual learning benchmark for multilingual asr," *arXiv preprint arXiv:2310.16931*, 2023.

[213] E. Verwimp, K. Yang, S. Parisot, L. Hong, S. McDonagh, E. Pérez-Pellitero, M. De Lange, and T. Tuytelaars, "Clad: A realistic continual learning benchmark for autonomous driving," *Neural Networks*, vol. 161, pp. 659–669, 2023.

[214] M. M. Derakhshani, I. Najdenkoska, T. van Sonsbeek, X. Zhen, D. Mahapatra, M. Worring, and C. G. Snoek, "Lifelonger: A benchmark for continual disease classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 314–324.

[215] M. Wołczyk, M. Zając, R. Pascanu, Ł. Kuciński, and P. Miłoś, "Continual world: A robotic benchmark for continual reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28 496–28 510, 2021.

[216] J. Ko, S. Kang, and K. Shin, "Begin: Extensive benchmark scenarios and an easy-to-use framework for graph continual learning," *arXiv preprint arXiv:2211.14568*, 2022.

[217] K. Faber, D. Zurek, M. Pietron, N. Japkowicz, A. Vergari, and R. Corizzo, "From mnist to imagenet and back: Benchmarking continual curriculum learning," *arXiv preprint arXiv:2303.11076*, 2023.

[218] Z. Lin, J. Shi, D. Pathak, and D. Ramanan, "The clear benchmark: Continual learning on real-world imagery," in *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

[219] Z. Cai, O. Sener, and V. Koltun, "Online continual learning with natural distribution shifts: An empirical study with visual data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8281–8290.

[220] J. Wang, X. Wang, Y. Shang-Guan, and A. Gupta, "Wanderlust: Online continual object detection in the real world," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 829–10 838.

[221] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, 2017.

[222] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[223] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "Podnet: Pooled outputs distillation for small-tasks incremental learning," in *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*. Springer, 2020, pp. 86–102.

[224] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Hkf2_sC5FX

[225] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 532–547.

[226] Y. Wu, T. Shi, K. Sharma, C. W. Seah, and S. Zhang, "Online continual knowledge learning for language models," *arXiv preprint arXiv:2311.09632*, 2023.

[227] M. Davari, N. Asadi, S. Mudur, R. Aljundi, and E. Belilovsky, "Probing representation forgetting in supervised and unsupervised continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 712–16 721.

[228] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[229] H. A. Al Kader Hammoud, A. Prabhu, S.-N. Lim, P. H. Torr, A. Bibi, and B. Ghanem, "Rapid adaptation in online continual learning: Are we evaluating it right?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 852–18 861.

[230] D. Yogatama, C. d. M. d'Autume, J. Connor, T. Kocisky, M. Chrzanowski, L. Kong, A. Lazaridou, W. Ling, L. Yu, C. Dyer *et al.*, "Learning and evaluating general linguistic intelligence," *arXiv preprint arXiv:1901.11373*, 2019.

[231] L. Blier and Y. Ollivier, "The description length of deep learning models," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[232] B. Liu, S. Mazumder, E. Robertson, and S. Grigsby, "Ai autonomy: Self-initiated open-world continual learning and adaptation," *AI Magazine*, 2023.

[233] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[234] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[235] O. Goldreich, "Secure multi-party computation," *Manuscript. Preliminary version*, vol. 78, no. 110, pp. 1–108, 1998.