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

YUTAO YANG, JIE ZHOU*, XUANWEN DING, TIANYU HUAI, SHUNYU LIU, QIN CHEN, LIANG HE, and YUAN XIE, School of Computer Science and Technology, East China Normal University, China

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.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Scene understanding**; *Cognitive robotics*; *Cognitive science*; **Intelligent agents**.

Additional Key Words and Phrases: Continual Learning, Foundation Language Models, Pre-trained Language Models, Large Language Models, Vision-Language Models, Survey

ACM Reference Format:

Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Liang He, and Yuan Xie. 2018. Recent Advances of Foundation Language Models-based Continual Learning: A Survey. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 40 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advancements in foundation language models (LMs) have set new benchmarks in both natural language processing (NLP) [136, 226, 232] and computer vision (CV) [188]. Foundation LMs encompass three primary categories: Pre-trained Language Models (PLMs) [136], Large Language Models (LLMs) [226], and Vision-Language Models (VLMs) [42]. PLMs such as BERT [88], RoBERTa [120], and BART [102] focus on text-based tasks and are crucial for

*Corresponding authors.

Authors' Contact Information: Yutao Yang; Jie Zhou, jzhou@cs.ecnu.edu.cn; Xuanwen Ding; Tianyu Huai; Shunyu Liu; Qin Chen; Liang He; Yuan Xie, School of Computer Science and Technology, East China Normal University, Shanghai, China.

Manuscript submitted to ACM

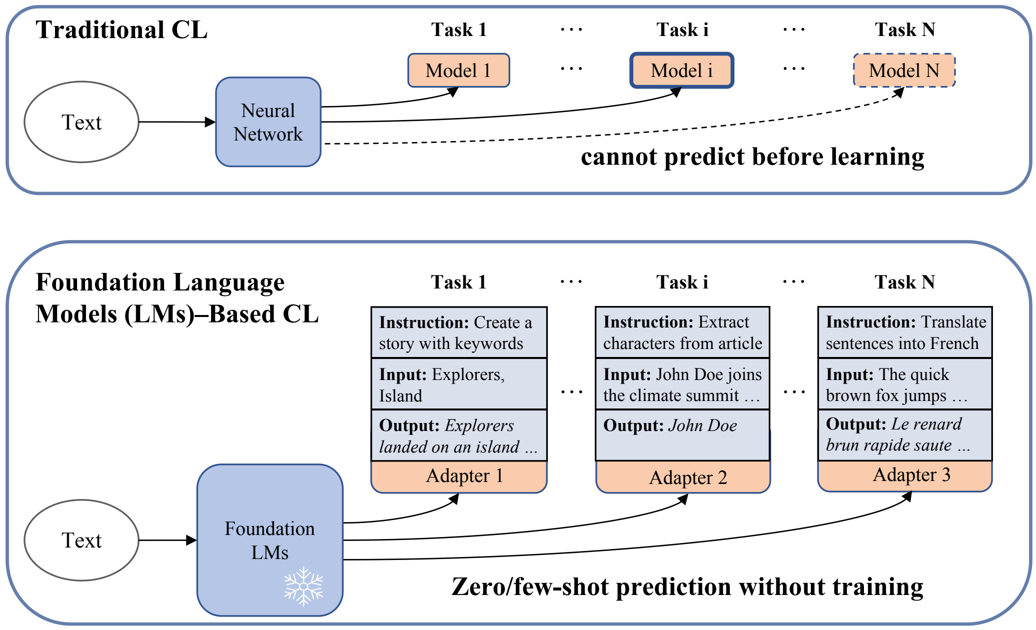


Fig. 1. Comparison between traditional CL and Foundation language models (LMs)-Based CL.

understanding and generating language by leveraging tasks like masked language modeling during pre-training. LLMs, including models like GPT-4 [1] and LLaMA [173], 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 [106], CLIP [154], LLaVA [113] and DALL-E [156], 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.

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 integrate new data or concepts without undergoing a retraining process. A significant challenge associated with training on a sequence of tasks is "catastrophic forgetting" [92], 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) [175, 186], also referred to as lifelong learning [145] or incremental learning [230], 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 [34]. This

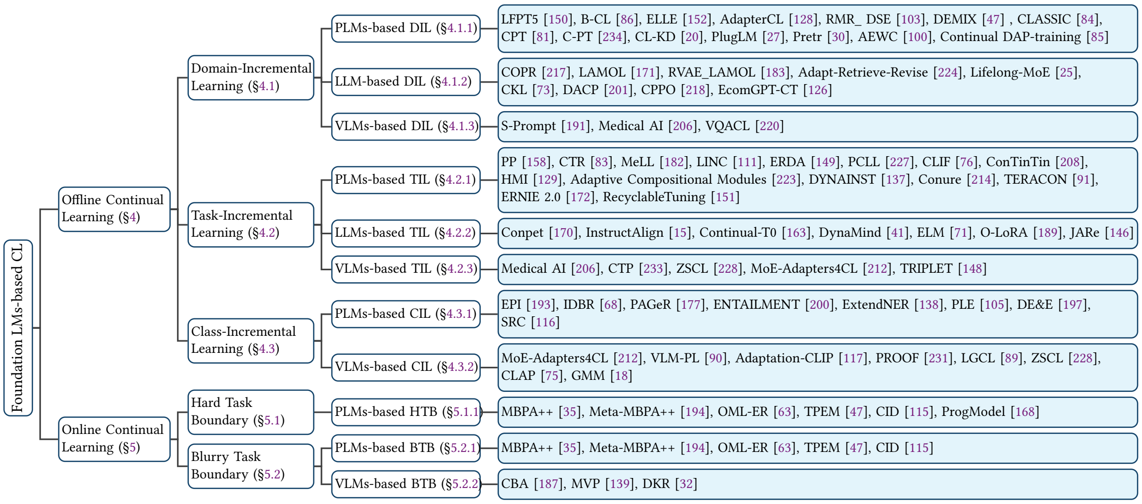


Fig. 2. Taxonomy of foundation language models for continual learning.

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 [30, 99, 134]. 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 [84]. Similarly, in dialogue generation, the novel technologies assist models in refining and expanding their conversational capabilities through ongoing interactions [164]. In text classification, continual learning facilitates the incorporation of new categories and adjustments to shifts in text distributions without the need for complete retraining [158]. 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 [148, 220]. 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 [59], such as prompt tuning [119] and adapters [140], to update parameters without comprehensive retraining. Third, the foundation LMs possess the capability to follow instructions through instructional learning [39, 144], 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.

The paper is organized as follows. In Section 2, we review the mainly related surveys about continual learning. Then, we introduce the base settings and learning modes of continual learning in Section 3, 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 4. In Section 5, we focus on online continual learning, including hard task boundary and blurry task boundary settings. The typical datasets and metrics are provided in Section 6 and 7. Finally, we analyze the challenge and further work in Section 8 and give the conclusion in Section 9.

2 RELATED SURVEYS

2.1 Continual Learning

Early examinations of Continual Learning (CL) have provided broad coverage, as observed in surveys such as Parisi et al. [145]. Recently, Wang et al. [186] conduct 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 [7, 131, 230] and replay-based approaches [60], reflecting the increasing granularity of research interests within the CL domain.

2.2 Continual Learning for Computer Vision

In the realm of computer vision, De et al. [34] 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 a comprehensive experimental analysis, comparing the efficacy of 11 methodologies across three benchmarks. Qu et al. [153] 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. [130] 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. [131] 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. [7] 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 an evaluation framework for detailed analysis.

2.3 Continual Learning for NLP

Biesialska et al. [8] 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. [82] 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.

2.4 Continual Learning for Other Domains

Recent surveys, surveys like [101, 167, 219] investigate the advancements in incremental learning for neural recommendation systems and continual learning (CL) in robotics, respectively. Zhang et al. [219] 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. [167] 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. [101] explore CL within this domain, defining CL as a paradigm where both data distribution and learning

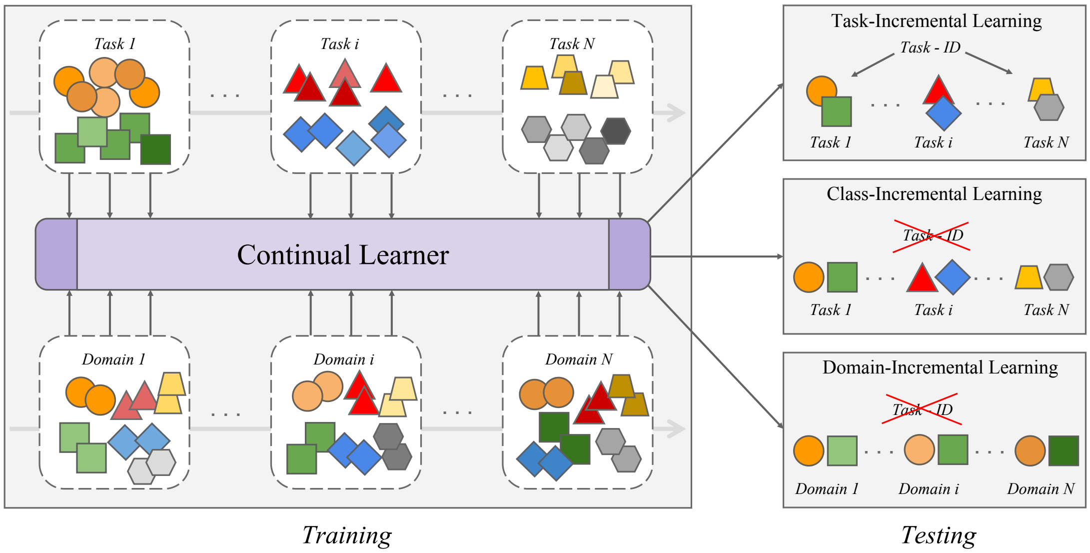


Fig. 3. The setting of different offline continual learning tasks, including task-incremental learning, class-incremental learning and domain-incremental learning. The samples with different classes (domains) are marked with various shapes (colors).

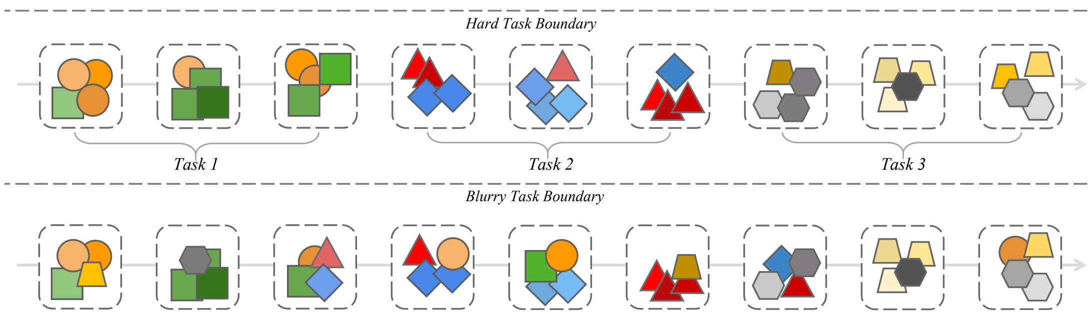


Fig. 4. The setting of different online continual learning tasks, including hard task boundary arriving and blurry task boundary arriving. The samples with different classes (domains) are marked with various shapes (colors).

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 and online CL based on PLMs, LLMs, and VLMs.

3 SETTINGS AND LEARNING MODES OF CL

3.1 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.

3.2 Typical Scenarios

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

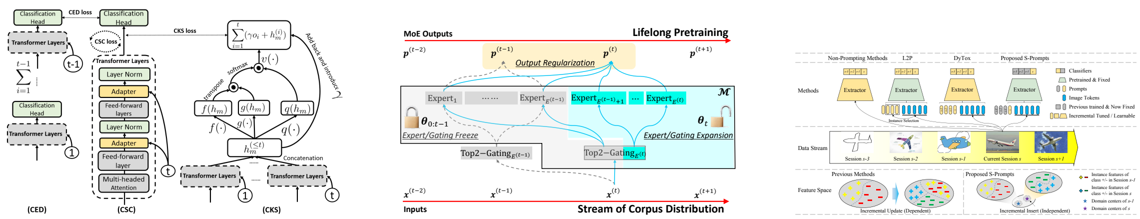
- Domain-Incremental Learning (DIL): The model aims to process diverse data distributions. Specifically, in DIL, while the data distributions $p(X_t)$ in task t and $p(X_{t'})$ in task t' are different, their task types and class labels remain consistent. The task identities (task IDs) are not required.
- Task-Incremental Learning (TIL): 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 the training and testing phases.
- Class-Incremental Learning (CIL): The model is designed to continually learn new class 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.

3.2.2 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 next task, ensuring no overlap of data between tasks.
- Blurry Task Boundary: The distinction between tasks is less clear, similar to 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 [152, 171], elastic weight consolidation (EWC) [92], and progressive neural networks [83, 86], have emerged to address this issue. Each method comes with its unique strengths and weaknesses upon the task arrival configuration.



(a) CLASSIC (b) Lifelong-MoE (c) S-Prompts

Fig. 5. Frameworks in DIL: CLASSIC (PLM-based) [84], Lifelong-MoE (LLM-based) [25], S-Prompts (VLM-based) [191].

4 OFFLINE CONTINUAL LEARNING

4.1 Domain-Incremental Learning

4.1.1 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. [103] present a regularization-centric framework for Lifelong Learning (LLL) termed RMR-DSE, specifically tailored for sequential operation across multiple domains. 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. Castellucci et al. [20] propose a Knowledge Distillation-based Continual Learning (CL-KD) approach with 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. [100] introduce a domain-agnostic framework that 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 algorithm is applied to adjust the loss function, ensuring the balance between acquiring new knowledge and retaining previously learned information.

Gururangan et al. [57] 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 [27] is a 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 (DAP-training) [85] 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. 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. [30] 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 [64, 147] and p-tuning [119] are used for domain-incremental CL [84, 86, 128, 234].

The adapter architecture incorporates a skip-connection to minimize the number of parameters. A notable exemplar of this approach is AdapterCL [128], which employs residual adapters tailored specifically 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. [86] introduce B-CL model to tackle critical challenges in CL for Aspect-Based Sentiment Classification. B-CL integrates continual learning adapters within capsule network architectures. Aiming at mitigating catastrophic forgetting during fine-tuning, the CLASSIC model [84] presents an innovative solution by deploying adapters to tap into BERT's capabilities (Figure 5a). 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) [81] introduces two continual learning plug-in modules, termed CL-plugins, embedded within each transformer layer of RoBERTa.

Prompt tuning [119], 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. [150] 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 representative 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. [234] 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. [152] 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.

4.1.2 LLMs-based DIL.

Traditional Methods. In many practical scenarios, retraining Language Models (LMs) is challenging due to resource constraints and data privacy concerns. Zhang et al. [218] 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. [217] propose Continual Optimal Policy Regularization (COPR), which calculates the optimal policy distribution without the partition function and uses the previous optimal policy to regularize the current policy. Sun et al. [171] 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. [183] 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.

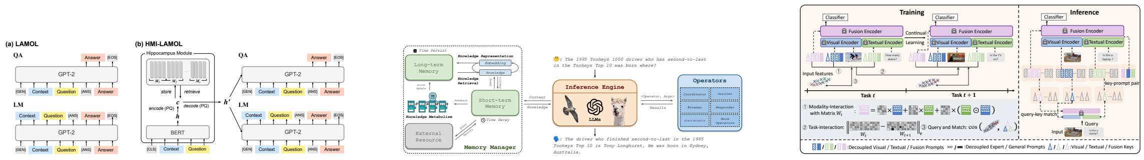
To reduce hallucinations in specialized domains such as the Chinese legal domain, Zhang et al. [224] propose a novel domain adaptation framework, named Adapt-Retrieve-Revise (ARR). It consists of three steps: adapting a 7-billion-parameter language model for initial responses, retrieving corroborative evidence from an external knowledge base, and integrating these to refine the final response with GPT-4. Gogoulou et al. [48] 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. [73]. To assess CKL approaches, the authors establish a benchmark and metric measuring knowledge retention, updating, and acquisition.

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 [26, 126, 201]. Among these studies, Cheng et al. [26] 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. [201] 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. [126] 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. [25] present an innovative Lifelong Learning framework, termed Lifelong-MoE, which leverages a Mixture-of-Experts (MoE) architecture (Figure 5b). This architecture enhances the model's capacity by incorporating new experts, where previously trained experts and gating mechanisms are frozen.

4.1.3 VLMs-based DIL. Vision-language models (VLMs) have demonstrated superiority in domain-incremental learning contexts. Yi et al. [206] 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, the prompt text is utilized to master the pre-trained knowledge embedded within VLMs. Aiming to independently learn prompts across disparate domains by using pre-trained VLMs, S-Prompt [191] is devised (Figure 5c). 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.



(a) HMI (b) DynaMind (c) TRIPLET

Fig. 6. Frameworks in TIL: HMI (PLM-based) [129], DynaMind (LLM-based) [41], TRIPLET (VLM-based) [148].

In the domain of visual question answering, Zhang et al. [220] 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.

4.2 Task-Incremental Learning

4.2.1 PLMs-based TIL.

Traditional Methods. Drawing inspiration from neurobiological mechanisms, Maekawa et al. [129] 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. [149] 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. [184] 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 [214] 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 [91] 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. [83] 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. [182] introduce a novel methodology, MeLL, which utilizes a BERT-based text encoder to generate 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.

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 [110, 111, 132]. 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 **??**). 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. [172] 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, RecyclableTuning [151] 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. [223] propose a pioneering strategy that 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. [76] 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. [227] 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. [158] 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 [208] 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 the 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. [137]. 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 Instruction 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.

4.2.2 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. [190] and Peng et al. [146]. 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. [41], emerges as a pioneering framework that 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. [125] 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. [190] 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. Peng et al. [146] 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) [170] 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 [71] 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. [189] 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. [163] 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, 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 [15] 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.

4.2.3 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 [233] employs topology preservation and momentum contrast to 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. [228] 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 [148] 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.

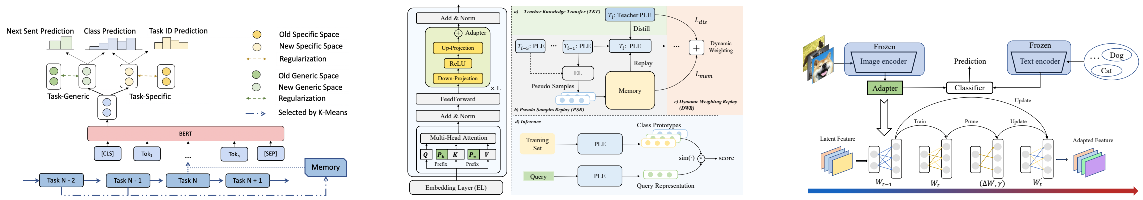
Moreover, COIN [23] introduces a new continuous instruction tuning benchmark designed to evaluate the performance of Multimodal Large Language Models (MLLMs) in the sequential instruction fine-tuning paradigm. CoIN includes 10 commonly used data sets covering 8 task categories, ensuring the diversity 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 [212] 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.

4.3 Class-Incremental Learning

4.3.1 PLMs-based CIL.

Traditional Methods. The study presented in [138] 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



(a) IDBR (b) PLE (c) Adaptation-CLIP

Fig. 7. Frameworks in CIL: IDBR (PLM-based) [68], PLE (PLM-based) [105], Adaptation-CLIP (VLM-based) [117].

allows the student to acquire the ability to recognize new entities while retaining knowledge of previously identified ones. Additionally, Liu et al. [116] 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. [68] 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. [177], 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. [105] propose a 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. [193], 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. [197] 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.

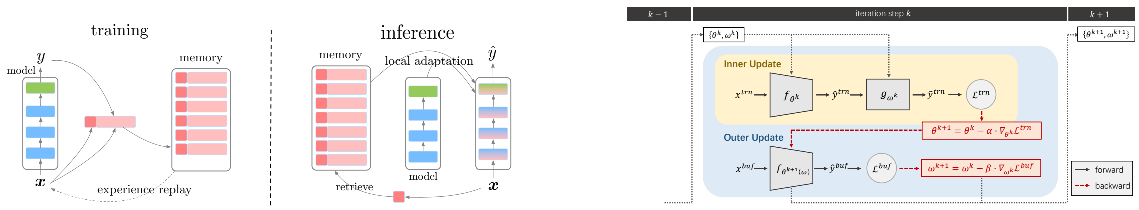
4.3.2 VLMs-based CIL.

Traditional Methods. Kim et al. [90] introduce 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. Cao et al. [18] 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.

PROOF [231] 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. [75] present 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. [117] 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. Khan et al. [89] 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,



(a) MBPA++ (b) CBA

Fig. 8. Frameworks in Online Continual Learning: MBPA++ (PLM-based HTB/BTB) [35], CBA (VLM-base BTB) [187].

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.

5 ONLINE CONTINUAL LEARNING

5.1 Hard Task Boundary

5.1.1 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. For example, Shen et al. [168] have implemented HTB in slot filling, Michieli et al. [135] have utilized it in audio classification, and Vander et al. [176] 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 [35, 194], replay-based methods [63] and regularization-based methods [115, 176]. MBPA++ [35] 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++ [194] 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. [115] 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.

5.2 Blurry Task Boundary

5.2.1 PLMs-based BTB. MBPA++ [35] and Meta-MBPA++ [194] can also exemplify models capable of adapting to environments with indistinct task boundaries. TPEM [47] employs a tripartite approach within an encoder-decoder framework, consisting of pruning, expanding, and masking techniques. Pruning helps preserve essential 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) [74] and *a neuromodulatory meta-learning algorithm* (ANML) [6] were intended to continuously learn new sequences of tasks during the testing phase. Holla et al. [63] 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.

5.2.2 VLMs-based BTB.

Traditional Methods. Cui et al. [32] propose the Dynamic Knowledge Rectification (DKR) framework, designed to mitigate the propagation of incorrect information in foundation LMs. 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 accuracy.

Parameter-Efficient Tuning Methods. Wang et al. [187] introduce the Continual Bias Adaptor (CBA) (Figure 8b), a novel method designed to enhance the efficacy of online 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. [139], 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.

Online Learning for CV. Recent advancements in class-incremental online learning predominantly address computer vision tasks, encapsulating various methodologies such as regularization-based [5, 52, 54, 207], replay-based [3, 14, 107, 169, 222], distillation-based [46, 61, 95], and gradient-based approaches [24, 55, 77, 143, 210]. Among these, Koh et al. [94] 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. [53] 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. [195] and Chrysalis et al. [29]. Additionally, Semola et al. [166] 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.

6 DATASETS

6.1 Offline Datasets for NLP

6.1.1 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 [221], including AG News, Amazon Reviews, Yelp Reviews, DBpedia, and Yahoo Answers [171]. Particularly, the AG News dataset has 4 classes for news

Table 1. The statistics information of the existing CL datasets. #D/T/C means the number of domains/tasks/classes, respectively.

Datasets	#Train	#Val	#Test	#Total	CL Settings	NLP Problems	Language	#D/T/C
Offline								
Progressive Prompts [158]	-	-	-	-	TIL	Sentiment analysis, topic classification, boolean QA, QA, paraphrase detection, word sense disambiguation, natural language inference	English	15 tasks
MeLL [182]	1,430,880	173,781	118,240	1,722,901	TIL	Intent classification	English	1184 tasks
Continual-T0 [164]	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 [217]	-	-	-	-	TIL	QA tasks, summary task, positive file review generation task	English	3 tasks
Adaptive Compositional Modules [223]	50,725	-	27,944	78,669	TIL	Natural language generation, SQL query generation, Summarization and Task-oriented dialogue	English	8 tasks
CODETASKCL [204]	181,000	9,700	10,000	200,700	TIL	Code generation, code translation, code summarization, and code refinement	Hybrid	4 tasks
Lifelong Simple Questions [184]	-	-	-	-	TIL	single-relation questions	English	20 tasks
Lifelong FewRel [184]	-	-	-	-	TIL	few-shot relation detection	English	10 tasks
InstrDialog [225]	9,500	950	1,900	12,350	TIL	Dialogue state tracking, dialogue generation, intent identification	English	19 tasks
InstrDialog++ [225]	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 [208]	-	-	-	-	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 [214]	-	-	-	-	TIL	personalized recommendations and profile predictions	English	6 tasks
Movielens [214]	-	-	-	-	TIL	personalized recommendations and profile predictions	English	3 tasks
NAVER Shopping [91]	-	-	-	-	TIL	search query prediction tasks	English	6 tasks
TRACE [190]	40,000	-	16,000	56,000	TIL	Domain-specific task, Multi-lingual task, Code completion task, Mathematical reasoning task	Hybrid	8 tasks
ABSC [86]	3,452	150	1,120	4,722	DIL	Aspect-based sentiment classification	English	19 Domains
DecaNLP [171]	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 [171]	115,000	-	7,600	122,600	DIL	News classification, sentiment analysis, Wikipedia article classification, and question-and-answer categorization	English	5 domains
RVAE_LAMOL [183]	15,870	-	5,668	21,538	DIL	Oriented dialogue of the restaurant reservation task, semantic role labeling, sentiment classification	English	3 domains
COPR [217]	-	-	-	-	DIL	-	English	18 domains
SGD [234]	38,745	5,210	11,349	40,287	DIL	Dialogue state tracking	English	19 Domains
CPT [81]	3,121,926	-	-	3,121,926	DIL	Domain-adaptive pre-training task	English	4 Domains
CKL[73]	-	-	-	30,372	DIL	Domain-adaptive pre-training task	English	3 Domains
ELLE [152]	-	-	-	-	DIL	Domain-adaptive pre-training task	English	5 Domains
Domain-incremental Paper Stream [78]	-	-	-	-	DIL	Relation extraction and named entity recognition	English	4 domains
Chronologically-ordered Tweet Stream [78]	-	-	-	-	DIL	multi-label hashtag prediction and single-label emoji prediction	English	4 domains
AdapterCL [128]	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 [197]	28,982	-	12,089	41,071	CIL	Text classification	English	3 tasks
EPI [193]	12,840	3,524	6,917	23,281	CIL	Text classification, topic classification	English	13 Classes
PAGeR [177]	59,754	7,115	15,304	82,173	CIL	Intent classification	English	355 Classes
PLE [105]	4,669	4,650	31,642	40,961	CIL	Intent classification	English	477 Classes
CoNLL-03 [138]	23,326	5,902	5,613	34,841	CIL	Named Entity Recognition	English	4 Classes
OntoNotes [138]	107,169	16,815	10,464	134,448	CIL	Named Entity Recognition	English	6 Classes
Online								
Foundational text classification [35]	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++ [35]	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 [63]	-	-	-	-	Hard and Blurry	few-shot relation detection	English	10 tasks
Firehose [66]	-	-	-	110,000,000	Blurry	Personalized online language learning	English	1 tasks
TemporalWiki [72]	-	-	-	-	-	-	English	-

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. [158] 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 [181], SuperGLUE benchmark [180], and the IMDB dataset [127]. 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. DE&E [197] uses three common text classification data sets with different characteristics-News-groups, BBC News, and Consumer Finance Complaints2. Such datasets can be used to evaluate the models on tasks with different difficulty levels.

The datasets introduced in [193] 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 [221] 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) [96] and 20 Newsgroups [97], 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.

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 [177], aims to tackle the lifelong intent detection problem by combining three public intent classification datasets (CLINC150 [98], HWU64 [118], BANKING77 [19]), one text classification dataset (Stackoverflow S20 [203]), and two public multidomain dialog intent detection datasets (SGD [157], MWOZ [12]). Moreover, FewRel [58] 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 [105] consolidates nine well-regarded intent detection datasets, including CLINC150 [98] and HWU64 [118], 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 [182] 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 [31], TOP semantic parsing [56], and Facebook's Multilingual Task Oriented Dataset [162] into 90 tasks with overlapping label sets, amounting to over ten thousand samples. Hotline-EUIC is derived from an e-commerce dialogue system [104] and the hotline audios are transcribed to text by a high-accuracy industrial Automatic Speech Recognition (ASR) system.

Fine-grained Sentiment Analysis. Ke et al. [86] developed a task incremental learning dataset for 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. [67], which features consumer reviews for five different products; the Liu3Domains dataset by Liu [114], comprising reviews pertaining to three products; the Ding9Domains dataset by Ding et al. [38], which includes reviews of nine varied products; and the SemEval14 dataset, which is focused on reviews of two specific products—laptops and restaurants.

6.1.2 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 [164] focuses on English language generation tasks, including text simplification and empathetic dialogue generation, among others [9, 17]. 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. [125] conduct an analysis of catastrophic forgetting on Bloomz [161] using Continual T0 datasets.

The dataset, introduced in LAMOL [171], integrates elements from both DecaNLP [133] and the foundational text classification benchmark [221]. 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 task. Moreover, the dataset devised in RVAE_LAMOL [183], employs three tasks from DecaNLP: the English Wizard of Oz (WOZ) for goal-oriented dialogue, 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 [217] 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 [4], the summarization task based on the Reddit TL, DR dataset with human feedback [179], and the positive film review generation task using the IMDB dataset [127]. Meanwhile, the DIL framework requires the model to adapt to three distinct segments from the SHP dataset, as described by Ethayarajh et al. [44].

The dataset described in Adaptive Compositional Modules [223] explores sequence generation and categorizes tasks into "similar" and "dissimilar" groups based on their characteristics. Tasks classified as similar, including E2ENLG [142] and four domains (restaurant, hotel, TV, laptop) from RNNLG [196], 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) [229], CNN/DailyMail (news article summarization) [165], and MultiWOZ (semantic state sequence generation) [12] exhibit significant distributional shifts from previously encountered tasks. The CODETASKCL dataset, explored by Yadav et al. [204], encompasses a diverse array of code-centric tasks, including code generation [70], summarization [69], translation [123], and refinement [174] across various programming languages. This dataset significantly enhances the breadth of language processing applications within technical fields.

6.1.3 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 [138], exemplifies a continual learning approach to Named Entity Recognition (NER). This dataset amalgamates the CoNLL-03 English NER [160] and OntoNotes [65], covering a broad spectrum of entity types and sources. 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) [157] dataset, utilized in C-PT [234], 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 [184] is crafted for the task of relation extraction. It merges elements from the SimpleQuestions [11] and FewRel [58] to form a lifelong learning benchmark that confronts the challenges of relation detection in a few-shot context.

6.1.4 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 [81], 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 [202], AI and ACL Papers [121], and AGNews articles [221]. 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 [81], which evaluate domain-specific adaptability and incremental learning, the CKL benchmark [73] 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, UPDATEDLAMA, and NEWLAMA, which are crafted to probe specific types of knowledge that an LM may encounter in its learning trajectory.

Whereas the aforementioned two datasets assess more controlled dimensions of knowledge integration and retention, the dataset introduced in ELLE [152] 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) [235], NEWS ARTICLES (NS) [215], AMAZON REVIEWS (REV) [62], BIOMEDICAL PAPERS (BIO) [121] and COMPUTER SCIENCE PAPERS (CS) [121]. 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. [78] construct 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.

6.1.5 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 adaptability of models. A notable example is the dataset introduced in AdapterCL [128], which is tailored for task-oriented dialogue systems. This dataset incorporates four task-oriented datasets: TaskMaster 2019 (TM19) [13], TaskMaster 2020 (TM20) [13], Schema Guided Dialogue (SGD) [157], and MultiWoZ [12]. 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.

Continual Instruction Tuning Benchmark (CITB) [225] extends the concept of continual learning by focusing on instruction-based NLP tasks. Built on the comprehensive SuperNI [192] 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 ConTinTin [208] 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 a 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.

The dataset, used in Conure [214], consists of Tencent TL (TTL) [213] 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. [91] 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. [190], 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.

6.2 Online Datasets for NLP

6.2.1 Datasets for Classification. The foundational text classification benchmark, as introduced by Zhang et al. [221] 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++ [35] and OML-ER [63]

6.2.2 Datasets for Generation. The dataset, used in MBPA++ [35], comprises three distinct question-answering collections: SQuAD 1.1 [155], TriviaQA [79], and QuAC [28]. 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.

6.2.3 Datasets for Information Extraction. The lifelong relation extraction benchmark, used in OML-ER [63], is structured by Wang et al. [184] based on FewRel. Unlike the original application by Wang et al., the benchmark in OML-ER is adapted for online continuous learning scenarios.

6.2.4 Datasets for Other Tasks. Hu et al. [66] compile 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. TemporalWiki [72] addresses temporal misalignment 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.

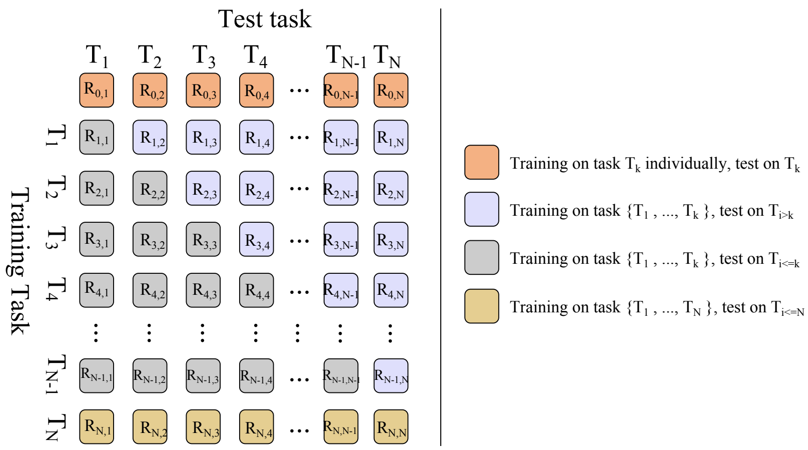


Fig. 9. Illustration of calculating metrics.

6.3 Offline CL Datasets for Multi-modal Tasks

The P9D dataset [233] 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. Qian et al. [148] introduce two novel benchmarks for continual learning, namely CL-TDIUC and CL-VQA2.0, which are derived from the TDIUC [80] and VQA2.0 [50], respectively. These benchmarks are categorized into three scenarios: the Continual Vision Scenario, which deals with new visual scenes; the Continual Language Scenario, focusing on new questions in existing scenes; and the Continual Vision-Language Scenario, addressing changes in both questions and visuals. DKR [32] comprises five benchmark datasets: MS-COCO Caption (MS-COCO) [108], Flickr30K [211], IAPR TC-12 [51], ECommerce-T2I (EC) [205], and RSICD [124]. 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. [141], 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.

6.4 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 [36, 176], autonomous driving [178], disease classification [37], reinforcement learning [198], graph data [93], and computer vision [45, 109]. Furthermore, some online CL benchmarks are also proposed for computer version tasks, such as continual visual learning [16], continual object detection [185].

7 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 9) 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 training on task T_i.

7.1 Overall Performance.

The metric termed "Last" [122, 228] 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. [228] 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. [159] and Douillard et al. [40], 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.

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

Moreover, Chaudhry et al. [22] devise a metric known as Learning Curve Area (LCA), which quantifies the speed of learning in a model. Qin et al. [152] 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^+).

7.2 Memory Stability.

Memory stability is typically evaluated by backward transfer (BWT) [122] and forgetting measure (FM) [21].

Backward Transfer (BWT) emerges as a pivotal concept extensively documented in the literature, notably by Lopez et al. [122] and Wu et al. [199]. BWT measures the performance degradation on previously mastered tasks after the model is trained on new tasks. This performance degradation phenomenon is often referred to as “forgetting".

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

Additionally, Chaudhry et al. [21] introduce the Forgetting Measure (FM), a metric designed to quantify the extent of forgetting a model experiences for a specific task. A lower FM indicates better retention of previous tasks. Davari et al. [33] 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. Kemker et al. [87] introduce three metrics, where 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. Additionally, researchers [95] devise a novel metric, termed the Knowledge Loss Ratio (KLR), quantifies knowledge degradation using principles from information theory.

7.3 Learning Plasticity.

Evaluating learning plasticity can be effectively accomplished through two key metrics: forward transfer (FWT) [122] and intransigence measure (IM) [21].

Forward Transfer (FWT) [122] assesses the beneficial effects on the performance of subsequent tasks following a model’s training on prior tasks.

$$FWT = \frac{1}{N-1}\sum_{i=2}^{N} R_{i-1,i} - R_{0,i} \tag{6}$$

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. [21], 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. Moreover, Koh et al. [95] introduce novel metrics, known as Knowledge Gain Ratio (KGR), which quantifies the capacity to acquire new knowledge through the calculation of knowledge gain.

7.4 Metrics for Continual Pre-training.

CKL [73] 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.

7.5 Online CL-Specific Metrics.

Near-future accuracy (NFA) [2] 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. Yogatama et al. [209] proposed a novel online codelength, inspired by prequential encoding [10], to quantify how quickly an existing model can adapt to a new task.

8 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. [112] 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 [111]. 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 multi-modal 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 [43], federated learning [216], and secure multi-party computation [49] 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 6, 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.

9 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.

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Hasan Abed Al Kader Hammoud, Ameya Prabhu, Ser-Nam Lim, Philip HS Torr, Adel Bibi, and Bernard Ghanem. 2023. Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?. In *ICCV*. 18852–18861.

[3] Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. 2019. Online Continual Learning with Maximal Interfered Retrieval. In *NeurIPS*, Vol. 32.

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[5] Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. 2022. Online continual learning on a contaminated data stream with blurry task boundaries. In *CVPR*. 9275–9284.

[6] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. 2020. Learning to continually learn. *arXiv preprint arXiv:2002.09571* (2020).

[7] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. 2021. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks* 135 (2021), 38–54.

[8] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. 2020. Continual Lifelong Learning in Natural Language Processing: A Survey. In *COLING*. 6523–6541.

[9] Raad Bin Tareaf. 2017. Tweets Dataset - Top 20 most followed users in Twitter social platform.

[10] Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. *NeurIPS* 31 (2018).

[11] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075* (2015).

[12] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278* (2018).

[13] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358* (2019).

[14] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. 2021. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025* (2021).

[15] Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-Low Resource Language Alignment via Continual Crosslingual Instruction Tuning. In *the First Workshop in South East Asian Language Processing*. 55–78.

[16] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. 2021. Online continual learning with natural distribution shifts: An empirical study with visual data. In *ICCV*. 8281–8290.

[17] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *NeurIPS* 31 (2018).

[18] Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, and Ming-Ming Cheng. 2024. Generative Multi-modal Models are Good Class-Incremental Learners. *arXiv preprint arXiv:2403.18383* (2024).

[19] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807* (2020).

[20] Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2021. Learning to Solve NLP Tasks in an Incremental Number of Languages. In *ACL*. 837–847.

[21] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*. 532–547.

[22] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019. Efficient Lifelong Learning with A-GEM. In *ICLR*.

[23] Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. 2024. CoIN: A Benchmark of Continual Instruction tuNing for Multimodel Large Language Model. *arXiv preprint arXiv:2403.08350* (2024).

[24] Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. 2020. Mitigating forgetting in online continual learning via instance-aware parameterization. *NeurIPS* 33 (2020), 17466–17477.

[25] Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023. Lifelong language pretraining with distribution-specialized experts. In *ICML*. PMLR, 5383–5395.

[26] Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530* (2023).

[27] Xin Cheng, Yankai Lin, Xiuying Chen, Dongyan Zhao, and Rui Yan. 2023. Decouple knowledge from paramters for plug-and-play language modeling. In *ACL*. 14288–14308.

[28] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *EMNLP*. 2174–2184.

[29] Aristotelis Chrysakis and Marie-Francine Moens. 2023. Online bias correction for task-free continual learning. *ICLR* (2023).

[30] Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357* (2022).

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

[32] Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. 2024. Continual Vision-Language Retrieval via Dynamic Knowledge Rectification. In *AAAI*, Vol. 38. 11704–11712.

[33] MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. 2022. Probing representation forgetting in supervised and unsupervised continual learning. In *CVPR*. 16712–16721.

[34] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *TPAMI* 44, 7 (2021), 3366–3385.

[35] Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *NeurIPS* 32 (2019).

[36] Luca Della Libera, Pooneh Mousavi, Salah Zaiem, Cem Subakan, and Mirco Ravanelli. 2023. CL-MASR: A Continual Learning Benchmark for Multilingual ASR. *arXiv preprint arXiv:2310.16931* (2023).

[37] Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, Marcel Worring, and Cees GM Snoek. 2022. Lifelonger: A benchmark for continual disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 314–324.

[38] Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM*. 231–240.

[39] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

[40] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*. 86–102.

[41] Mingzhe Du, Anh Tuan Luu, Bin Ji, and See-kiong Ng. 2023. From Static to Dynamic: A Continual Learning Framework for Large Language Models. *arXiv preprint arXiv:2310.14248* (2023).

[42] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A Survey of Vision-Language Pre-Trained Models. In *IJCAI*. 5436–5443.

[43] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. 1–12.

[44] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with V-usable information. In *ICML*. 5988–6008.

[45] Kamil Faber, Dominik Zurek, Marcin Pietron, Nathalie Japkowicz, Antonio Vergari, and Roberto Corizzo. 2023. From MNIST to ImageNet and Back: Benchmarking Continual Curriculum Learning. *arXiv preprint arXiv:2303.11076* (2023).

[46] Enrico Fini, Stéphane Lathuiliere, Enver Sangineto, Moin Nabi, and Elisa Ricci. 2020. Online continual learning under extreme memory constraints. In *ECCV 2020*. 720–735.

[47] Binzong Geng, Fajie Yuan, Qiancheng Xu, Ying Shen, Ruifeng Xu, and Min Yang. 2021. Continual Learning for Task-oriented Dialogue System with Iterative Network Pruning, Expanding and Masking. In *ACL*. 517–523.

[48] Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2023. A study of continual learning under language shift. *arXiv preprint arXiv:2311.01200* (2023).

[49] Oded Goldreich. 1998. Secure multi-party computation. *Manuscript. Preliminary version* 78, 110 (1998), 1–108.

[50] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*. 6904–6913.

[51] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, Vol. 2.

[52] Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. 2022. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. In *CVPR*. 7442–7451.

[53] Nuwan Gunasekara, Bernhard Pfahringer, Heitor Murilo Gomes, and Albert Bifet. 2023. Survey on Online Streaming Continual Learning. In *IJCAI*.

[54] Yiduo Guo, Bing Liu, and Dongyan Zhao. 2022. Online continual learning through mutual information maximization. In *ICML*. 8109–8126.

[55] Yiduo Guo, Bing Liu, and Dongyan Zhao. 2023. Dealing with Cross-Task Class Discrimination in Online Continual Learning. In *CVPR*. 11878–11887.

[56] Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. *arXiv preprint arXiv:1810.07942* (2018).

[57] Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2022. DEMix Layers: Disentangling Domains for Modular Language Modeling. In *NAACL*. 5557–5576.

[58] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147* (2018).

[59] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *arXiv preprint arXiv:2403.14608* (2024).

[60] Tyler L Hayes, Giri P Krishnan, Maxim Bazhenov, Hava T Siegelmann, Terrence J Sejnowski, and Christopher Kanan. 2021. Replay in deep learning: Current approaches and missing biological elements. *Neural computation* 33, 11 (2021), 2908–2950.

[61] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. 2020. Incremental learning in online scenario. In *CVPR*. 13926–13935.

[62] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*. 507–517.

[63] Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Meta-learning with sparse experience replay for lifelong language learning. *arXiv preprint arXiv:2009.04891* (2020).

[64] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *ICML*. 2790–2799.

[65] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *NAACL*. 57–60.

[66] Hexiang Hu, Ozan Sener, Fei Sha, and Vladlen Koltun. 2022. Drinking from a firehose: Continual learning with web-scale natural language. *TPAMI* 45, 5 (2022), 5684–5696.

[67] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *SIGKDD*. 168–177.

[68] Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. 2021. Continual Learning for Text Classification with Information Disentanglement Based Regularization. In *ACL*. 2736–2746.

[69] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436* (2019).

[70] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2018. Mapping language to code in programmatic context. *arXiv preprint arXiv:1808.09588* (2018).

[71] Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. In *ICML*. 14702–14729.

[72] Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models. In *EMNLP*. 6237–6250.

[73] Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards Continual Knowledge Learning of Language Models. In *ICLR*.

[74] Khurram Javed and Martha White. 2019. Meta-learning representations for continual learning. *NeurIPS* 32 (2019).

[75] Saurav Jha, Dong Gong, and Lina Yao. 2024. CLAP4CLIP: Continual Learning with Probabilistic Finetuning for Vision-Language Models. *arXiv preprint arXiv:2403.19137* (2024).

[76] Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. 2021. Learn Continually, Generalize Rapidly: Lifelong Knowledge Accumulation for Few-shot Learning. In *EMNLP*. 714–729.

[77] Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. 2021. Gradient-based editing of memory examples for online task-free continual learning. *NeurIPS* 34 (2021), 29193–29205.

[78] Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora. In *NAACL*. 4764–4780.

[79] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*. 1601–1611.

[80] Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *ICCV*. 1965–1973.

[81] Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual Training of Language Models for Few-Shot Learning. In *EMNLP*. 10205–10216.

[82] Zixuan Ke and Bing Liu. 2022. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701* (2022).

[83] Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, and Lei Shu. 2021. Achieving Forgetting Prevention and Knowledge Transfer in Continual Learning. In *NeurIPS*, Vol. 34. 22443–22456.

[84] Zixuan Ke, Bing Liu, Hu Xu, and Lei Shu. 2021. CLASSIC: Continual and Contrastive Learning of Aspect Sentiment Classification Tasks. In *EMNLP*.

[85] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual Pre-training of Language Models. In *ICLR*.

[86] Zixuan Ke and Hu Xu. 2021. Adapting BERT for Continual Learning of a Sequence of Aspect Sentiment Classification Tasks. In *NAACL*.

[87] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *AAAI*, Vol. 32.

[88] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[89] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, and Muhammad Zeshan Afzal. 2023. Introducing language guidance in prompt-based continual learning. In *ICCV*. 11463–11473.

[90] Junsu Kim, Yunhoe Ku, Jihyeon Kim, Junuk Cha, and Seungryul Baek. 2024. VLM-PL: Advanced Pseudo Labeling approach Class Incremental Object Detection with Vision-Language Model. *arXiv preprint arXiv:2403.05346* (2024).

[91] Sein Kim, Namkyeong Lee, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2023. Task Relation-aware Continual User Representation Learning. In *SIGKDD*. 1107–1119.

[92] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *the national academy of sciences* 114, 13 (2017), 3521–3526.

[93] Jihoon Ko, Shinhwan Kang, and Kijung Shin. 2022. BeGin: Extensive Benchmark Scenarios and An Easy-to-use Framework for Graph Continual Learning. *arXiv preprint arXiv:2211.14568* (2022).

[94] Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. 2021. Online continual learning on class incremental blurry task configuration with anytime inference. *arXiv preprint arXiv:2110.10031* (2021).

[95] Hyunseo Koh, Minhyuk Seo, Jihwan Bang, Hwanjun Song, Deokki Hong, Seulki Park, Jung-Woo Ha, and Jonghyun Choi. 2022. Online Boundary-Free Continual Learning by Scheduled Data Prior. In *ICLR*.

[96] Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *ICMLA*. 364–371.

[97] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings*. 331–339.

[98] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027* (2019).

[99] Kuan-Ying Lee, Yuanyi Zhong, and Yu-Xiong Wang. 2023. Do pre-trained models benefit equally in continual learning?. In *WCACV*. 6485–6493.

[100] Sungjin Lee. 2017. Toward continual learning for conversational agents. *arXiv preprint arXiv:1712.09943* (2017).

[101] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion* 58 (2020), 52–68.

[102] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

[103] Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. 2022. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *NAACL*. 5441–5454.

[104] Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017. Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *CIKM*. 2495–2498.

[105] Guodun Li, Yuchen Zhai, Qianglong Chen, Xing Gao, Ji Zhang, and Yin Zhang. 2022. Continual few-shot intent detection. In *COLING*. 333–343.

[106] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[107] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. 2023. PCR: Proxy-based Contrastive Replay for Online Class-Incremental Continual Learning. In *CVPR*. 24246–24255.

[108] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. 740–755.

[109] Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. 2021. The clear benchmark: Continual learning on real-world imagery. In *NeurIPS*.

[110] Bing Liu. 2020. Learning on the job: Online lifelong and continual learning. In *AAAI*, Vol. 34. 13544–13549.

[111] Bing Liu and Sahisnu Mazumder. 2021. Lifelong and continual learning dialogue systems: learning during conversation. In *AAAI*, Vol. 35. 15058–15063.

[112] Bing Liu, Sahisnu Mazumder, Eric Robertson, and Scott Grigsby. 2023. AI Autonomy: Self-initiated Open-world Continual Learning and Adaptation. *AI Magazine* (2023).

[113] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *NeurIPS* 36 (2024).

[114] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *IJCAI*.

[115] Qingbin Liu, Xiaoyan Yu, Shizhu He, Kang Liu, and Jun Zhao. 2021. Lifelong intent detection via multi-strategy rebalancing. *arXiv preprint arXiv:2108.04445* (2021).

[116] Tianlin Liu, Lyle Ungar, and João Sedoc. 2019. Continual Learning for Sentence Representations Using Conceptors. In *NAACL*. 3274–3279.

[117] Xialei Liu, Xusheng Cao, Haori Lu, Jia-wen Xiao, Andrew D Bagdanov, and Ming-Ming Cheng. 2023. Class Incremental Learning with Pre-trained Vision-Language Models. *arXiv preprint arXiv:2310.20348* (2023).

[118] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *International Workshop on Spoken Dialogue Systems*. 165–183.

[119] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *ACL*. 61–68.

[120] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[121] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *ACL*. 4969–4983.

[122] David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *NeurIPS* 30 (2017).

[123] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664* (2021).

[124] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing* 56, 4 (2017), 2183–2195.

[125] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. *arXiv e-prints* (2023).

[126] Shirong Ma, Shen Huang, Shulin Huang, Xiaobin Wang, Yangning Li, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. EcomGPT-CT: Continual pre-training of e-commerce large language models with semi-structured data. *arXiv preprint arXiv:2312.15696* (2023).

[127] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *ACL*. 142–150.

[128] Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, and Zhiguang Wang. 2020. Continual learning in task-oriented dialogue systems. *arXiv preprint arXiv:2012.15504* (2020).

[129] Aru Maekawa, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2023. Generative Replay Inspired by Hippocampal Memory Indexing for Continual Language Learning. In *EACL*. 930–942.

[130] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing* 469 (2022), 28–51.

[131] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. 2022. Class-incremental learning: survey and performance evaluation on image classification. *TPAMI* 45, 5 (2022), 5513–5533.

[132] Sahisnu Mazumder and Bing Liu. 2024. *Lifelong and Continual Learning Dialogue Systems.* Springer Nature.

[133] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. [n. d.]. The natural language decathlon: Multitask learning as question answering. arXiv 2018. *arXiv preprint arXiv:1806.08730* ([n. d.]).

[134] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. 2021. An Empirical Investigation of the Role of Pre-training in Lifelong Learning. (2021).

[135] Umberto Michieli, Pablo Peso Parada, and Mete Ozay. 2023. Online Continual Learning in Keyword Spotting for Low-Resource Devices via Pooling High-Order Temporal Statistics. *arXiv preprint arXiv:2307.12660* (2023).

[136] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.

[137] Jisoo Mok, Jaeyoung Do, Sungjin Lee, Tara Taghavi, Seunghak Yu, and Sungroh Yoon. 2023. Large-scale Lifelong Learning of In-context Instructions and How to Tackle It. In *ACL*. 12573–12589.

[138] Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko. 2021. Continual learning for named entity recognition. In *AAAI*, Vol. 35. 13570–13577.

[139] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. 2023. Online Class Incremental Learning on Stochastic Blurry Task Boundary via Mask and Visual Prompt Tuning. In *ICCV*. 11731–11741.

[140] Nandini Mundra, Sumanth Doddapaneni, Raj Dabre, Anoop Kunchukuttan, Ratish Puduppully, and Mitesh M Khapra. 2024. A Comprehensive Analysis of Adapter Efficiency. In *IKDD*. 136–154.

[141] Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. 2023. Continual vision-language representation learning with off-diagonal information. In *ICML*. 26129–26149.

[142] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254* (2017).

[143] Alex Ororbia, Ankur Mali, C Lee Giles, and Daniel Kifer. 2022. Lifelong neural predictive coding: Learning cumulatively online without forgetting. *NeurIPS* 35 (2022), 5867–5881.

[144] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS* 35 (2022), 27730–27744.

[145] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks* 113 (2019), 54–71.

[146] Bohao PENG, Zhuotao Tian, Shu Liu, Ming-Chang Yang, and Jiaya Jia. 2024. Scalable Language Model with Generalized Continual Learning. In *ICLR*.

[147] Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. In *EMNLP*. 149–160.

[148] Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. 2023. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In *ICCV*. 2953–2962.

[149] Chengwei Qin and Shafiq Joty. 2022. Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation. In *ACL*. 2776–2789.

[150] Chengwei Qin and Shafiq Joty. 2022. LFPT5: A Unified Framework for Lifelong Few-shot Language Learning Based on Prompt Tuning of T5. In *ICLR*.

[151] Yujia Qin, Cheng Qian, Xu Han, Yankai Lin, Huadong Wang, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Recyclable Tuning for Continual Pre-training. In *ACL*. 11403–11426.

[152] Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. ELLE: Efficient Lifelong Pre-training for Emerging Data. In *ACL*. 2789–2810.

[153] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. 2021. Recent advances of continual learning in computer vision: An overview. *arXiv preprint arXiv:2109.11369* (2021).

[154] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.

[155] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*. 2383–2392.

[156] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*. Pmlr, 8821–8831.

[157] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI*, Vol. 34. 8689–8696.

[158] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive Prompts: Continual Learning for Language Models. In *ICLR*.

[159] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental Classifier and Representation Learning. In *CVPR*.

[160] Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050* (2003).

[161] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[162] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327* (2018).

[163] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned Language Models are Continual Learners. In *EMNLP*. 6107–6122.

[164] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *EMNLP*. 6107–6122.

[165] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[166] Rudy Semola, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual-learning-as-a-service (claas): On-demand efficient adaptation of predictive models. *arXiv preprint arXiv:2206.06957* (2022).

[167] Khadija Shaheen, Muhammad Abdullah Hanif, Osman Hasan, and Muhammad Shafique. 2022. Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks. *Journal of Intelligent & Robotic Systems* 105, 1 (2022), 9.

[168] Yilin Shen, Xiangyu Zeng, and Hongxia Jin. 2019. A progressive model to enable continual learning for semantic slot filling. In *EMNLP*. 1279–1284.

[169] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. 2021. Online class-incremental continual learning with adversarial shapley value. In *AAAI*, Vol. 35. 9630–9638.

[170] Chenyang Song, Xu Han, Zheni Zeng, Kuai Li, Chen Chen, Zhiyuan Liu, Maosong Sun, and Tao Yang. 2023. ConPET: Continual Parameter-Efficient Tuning for Large Language Models. *arXiv preprint arXiv:2309.14763* (2023).

[171] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. LAMOL: LAnguage MOdeling for Lifelong Language Learning. In *ICLR*. OpenReview.net.

[172] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*, Vol. 34. 8968–8975.

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

[174] Michele Tufano, Cody Watson, Gabriele Bavota, Massimiliano Di Penta, Martin White, and Denys Poshyvanyk. 2019. An empirical study on learning bug-fixing patches in the wild via neural machine translation. *TOSEM* 28, 4 (2019), 1–29.

[175] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. 2022. Three types of incremental learning. *Nature Machine Intelligence* 4, 12 (2022), 1185–1197.

[176] Steven Vander Eeckt et al. 2023. Rehearsal-Free Online Continual Learning for Automatic Speech Recognition. *arXiv e-prints* (2023), arXiv–2306.

[177] Vaibhav Varshney, Mayur Patidar, Rajat Kumar, Lovekesh Vig, and Gautam Shroff. 2022. Prompt augmented generative replay via supervised contrastive learning for lifelong intent detection. In *NAACL*. 1113–1127.

[178] Eli Verwimp, Kuo Yang, Sarah Parisot, Lanqing Hong, Steven McDonagh, Eduardo Pérez-Pellitero, Matthias De Lange, and Tinne Tuytelaars. 2023. Clad: A realistic continual learning benchmark for autonomous driving. *Neural Networks* 161 (2023), 659–669.

[179] Michael Volske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to Learn Automatic Summarization. In *the Workshop on New Frontiers in Summarization*. 59–63.

[180] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *NeurIPS* 32 (2019).

[181] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP*. 353–355.

[182] Chengyu Wang, Haojie Pan, Yuan Liu, Kehan Chen, Minghui Qiu, Wei Zhou, Jun Huang, Haiqing Chen, Wei Lin, and Deng Cai. 2021. Mell: Large-scale extensible user intent classification for dialogue systems with meta lifelong learning. In *SIGKDD*. 3649–3659.

[183] Han Wang, Ruiliu Fu, Xuejun Zhang, and Jun Zhou. 2022. RVAE-LAMOL: Residual Variational Autoencoder to Enhance Lifelong Language Learning. In *IJCNN*. IEEE, 1–9.

[184] Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence Embedding Alignment for Lifelong Relation Extraction. In *NAACL*. 796–806.

[185] Jianren Wang, Xin Wang, Yue Shang-Guan, and Abhinav Gupta. 2021. Wanderlust: Online continual object detection in the real world. In *ICCV*. 10829–10838.

[186] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *TPAMI* (2024).

[187] Quanziang Wang, Renzhen Wang, Yichen Wu, Xixi Jia, and Deyu Meng. 2023. CBA: Improving Online Continual Learning via Continual Bias Adaptor. In *ICCV*. 19082–19092.

[188] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research* 20, 4 (2023), 447–482.

[189] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023. Orthogonal Subspace Learning for Language Model Continual Learning. In *EMNLP*. 10658–10671.

[190] Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. TRACE: A Comprehensive Benchmark for Continual Learning in Large Language Models.

[191] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. 2022. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *NeurIPS* 35 (2022), 5682–5695.

[192] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705* (2022).

[193] Zhicheng Wang, Yufang Liu, Tao Ji, Xiaoling Wang, Yuanbin Wu, Congcong Jiang, Ye Chao, Zhencong Han, Ling Wang, Xu Shao, et al. 2023. Rehearsal-free Continual Language Learning via Efficient Parameter Isolation. In *ACL*. 10933–10946.

[194] Zirui Wang, Sanket Vaibhav Mehta, Barnabas Poczos, and Jaime Carbonell. 2020. Efficient Meta Lifelong-Learning with Limited Memory. In *EMNLP*. 535–548.

[195] Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. 2023. Online Prototype Learning for Online Continual Learning. In *ICCV*. 18764–18774.

[196] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745* (2015).

[197] Mateusz Wójcik, Witold Kościukiewicz, Mateusz Baran, Tomasz Kajdanowicz, and Adam Gonczarek. 2023. Domain-Agnostic Neural Architecture for Class Incremental Continual Learning in Document Processing Platform. In *ACL*. 527–537.

[198] Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. 2021. Continual world: A robotic benchmark for continual reinforcement learning. *NeurIPS* 34 (2021), 28496–28510.

[199] Yuhao Wu, Tongjun Shi, Karthick Sharma, Chun Wei Seah, and Shuhao Zhang. 2023. Online Continual Knowledge Learning for Language Models. *arXiv preprint arXiv:2311.09632* (2023).

[200] Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. *arXiv preprint arXiv:2104.11882* (2021).

[201] Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *arXiv preprint arXiv:2311.08545* (2023).

[202] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232* (2019).

[203] Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks* 88 (2017), 22–31.

[204] Prateek Yadav, Qing Sun, Hantian Ding, Xiaopeng Li, Dejiao Zhang, Ming Tan, Xiaofei Ma, Parminder Bhatia, Ramesh Nallapati, Murali Krishna Ramanathan, et al. 2023. Exploring continual learning for code generation models. *arXiv preprint arXiv:2307.02435* (2023).

[205] An Yang, Junyang Lin, Rui Men, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Jiamang Wang, Yong Li, et al. 2021. M6-t: Exploring sparse expert models and beyond. *arXiv preprint arXiv:2105.15082* (2021).

[206] Huahui Yi, Ziyuan Qin, Qicheng Lao, Wei Xu, Zekun Jiang, Dequan Wang, Shaoting Zhang, and Kang Li. 2023. Towards General Purpose Medical AI: Continual Learning Medical Foundation Model. *arXiv preprint arXiv:2303.06580* (2023).

[207] Haiyan Yin, Ping Li, et al. 2021. Mitigating forgetting in online continual learning with neuron calibration. *NeurIPS* 34 (2021), 10260–10272.

[208] Wenpeng Yin, Jia Li, and Caiming Xiong. 2022. ConTinTin: Continual Learning from Task Instructions. In *ACL*. 3062–3072.

[209] Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373* (2019).

[210] Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. 2021. Online coreset selection for rehearsal-based continual learning. *arXiv preprint arXiv:2106.01085* (2021).

[211] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2 (2014), 67–78.

[212] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. 2024. Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters. In *CVPR*.

[213] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *SIGIR*. 1469–1478.

[214] Fajie Yuan, Guoxiao Zhang, Alexandros Karatzoglou, Joemon Jose, Beibei Kong, and Yudong Li. 2021. One Person, One Model, One World: Learning Continual User Representation without Forgetting. 696–705.

[215] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *NeurIPS* 32 (2019).

[216] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775.

[217] Han Zhang, Lin Gui, Yuanzhao Zhai, Hui Wang, Yu Lei, and Ruifeng Xu. 2023. Copf: Continual learning human preference through optimal policy fitting. *arXiv preprint arXiv:2310.15694* (2023).

[218] Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. 2024. CPPO: Continual Learning for Reinforcement Learning with Human Feedback. In *ICLR*.

[219] Peiyan Zhang and Sunghun Kim. 2023. A Survey on Incremental Update for Neural Recommender Systems. *arXiv preprint arXiv:2303.02851* (2023).

[220] Xi Zhang, Feifei Zhang, and Changsheng Xu. 2023. Vqacl: A novel visual question answering continual learning setting. In *CVPR*. 19102–19112.

[221] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *NeurIPS* 28 (2015).

[222] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. 2022. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. *NeurIPS* 35 (2022), 14771–14783.

[223] Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2022. Continual Sequence Generation with Adaptive Compositional Modules. In *ACL*. 3653–3667.

[224] Yating Zhang, Yexiang Wang, Fei Cheng, Sadao Kurohashi, et al. 2023. Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise. *arXiv preprint arXiv:2310.03328* (2023).

[225] Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. 2023. Citb: A benchmark for continual instruction tuning. *arXiv preprint arXiv:2310.14510* (2023).

[226] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[227] Yingxiu Zhao, Yinhe Zheng, Zhiliang Tian, Chang Gao, Jian Sun, and Nevin L. Zhang. 2022. Prompt Conditioned VAE: Enhancing Generative Replay for Lifelong Learning in Task-Oriented Dialogue. In *EMNLP*. 11153–11169.

[228] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. 2023. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *ICCV*. 19125–19136.

[229] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103* (2017).

[230] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2023. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648* (2023).

[231] Da-Wei Zhou, Yuanhan Zhang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2023. Learning without forgetting for vision-language models. *arXiv preprint arXiv:2305.19270* (2023).

[232] Jie Zhou, Pei Ke, Xipeng Qiu, Minlie Huang, and Junping Zhang. 2023. ChatGPT: potential, prospects, and limitations. *Frontiers of Information Technology & Electronic Engineering* (2023), 1–6.

[233] Hongguang Zhu, Yunchao Wei, Xiaodan Liang, Chunjie Zhang, and Yao Zhao. 2023. Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In *ICCV*. 22257–22267.

[234] Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual Prompt Tuning for Dialog State Tracking. In *ACL*. 1124–1137.

[235] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*. 19–27.

A DETAILS OF METRICS

A.1 Overall Performance.

Moreover, Chaudhry et al. [22] 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:

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

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

$$LCA_\beta = \frac{1}{\beta+1}\int_0^\beta Z_b \, db = \frac{1}{\beta+1}\sum_{b=0}^{\beta} Z_b \tag{8}$$

The Learning Curve Area (LCA) provides insights into model learning dynamics. LCA_0 measures the average 0-shot performance, similar to forward transfer ([122]). 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. [152] 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. [190].

A.2 Memory Stability.

Chaudhry et al. [21] 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:

$$f_j = \max_{l \in \{1,\dots,N-1\}} (R_{l,j} - R_{N,j}), \quad \forall j < N. \tag{9}$$

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:

$$FM = \frac{1}{N-1} \sum_{j=1}^{N-1} f_j \tag{10}$$

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. [33] 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: $W_i^* = \arg\min_{W_i} \mathcal{L}(W_i; f_{\theta_i}(X_i), Y_i)$, where $\mathcal{L}$, 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. [87] 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.

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

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

$$\Omega_{\text{all}} = \frac{1}{N-1} \sum_{i=2}^{N} \frac{\alpha_{\text{all},i}}{\alpha_{\text{ideal}}} \tag{13}$$

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

A.3 Learning Plasticity.

Intransigence measure (IM), as defined by Chaudhry et al. [21], 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.

A.4 Metrics for Continual Pre-training.

CKL [73] 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. Gap(T, D_a, D_b) is Score(T) of LM_a - 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.

A.5 Online CL-Specific Metrics.

Near-future accuracy (NFA) [2] 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. [209] proposed a novel online codelength ($\ell(D)$), inspired by prequential encoding [10], 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) [22], the concept of *online codelength* is associated with the area under the learning curve.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009