Bring Your Own Knowledge: A Survey of Methods for LLM Knowledge Expansion

Anonymous ACL submission

Abstract

Adapting large language models (LLMs) to new and diverse knowledge is essential for their lasting effectiveness in real-world applications. This survey provides an overview of state-of-the-art methods for expanding the knowledge of LLMs, focusing on integrating various knowledge types, including factual information, domain expertise, language proficiency, and user preferences. We explore techniques, such as continual learning, model editing, and retrieval-based explicit adaptation, while discussing challenges like knowledge consistency and scalability. Designed as a guide for researchers and practitioners, this survey sheds light on opportunities for advancing LLMs as adaptable and robust knowledge systems.

	Continual Learning (§4)	Model Editing (§5)	Retrieval (§6)
Knowledge Type			
Fact	✓	✓	✓
Domain	✓	✗	✓
Language	✓	✗	✗
Preference	✓	✓	✗
Applicability			
Large-scale data	✓	✗	✓
Precise control	✗	✓	✓
Computational cost	✗	✓	✓
Black-box applicable	✗	✗	✓

Table 1: We compare three key approaches for adapting LLMs — continual learning, model editing, and retrieval — based on their supported knowledge types and applicability across different criteria.

1 Introduction

As large language models (LLMs) are increasingly deployed in real-world applications, their ability to adapt to evolving knowledge becomes crucial for maintaining relevance and accuracy. However, LLMs are typically trained once and thus only have knowledge up to a certain cutoff date, limiting their ability to stay updated with new information. This survey provides a comprehensive overview of methods that enable LLMs to incorporate various types of new knowledge, including factual, domain-specific, language, and user preference knowledge. We survey adaptation strategies, including continual learning, model editing, and retrieval-based approaches, and aim at providing guidelines for researchers and practitioners.

To remain effective, LLMs require updates across multiple dimensions. Factual knowledge consists of general truths and real-time information, while domain knowledge pertains to specialized fields, such as medicine or law. Language knowledge enhances multilingual capabilities, and preference knowledge aligns model behavior with user expectations and values. Ensuring that LLMs can integrate updates across these dimensions is essential for their sustained utility.

Existing LLM adaptation methods differ in approach and application. Continual learning enables incremental updates to models' parametric knowledge, mitigating catastrophic forgetting (McCloskey and Cohen, 1989) while ensuring long-term performance. Model editing allows for precise modifications of learned knowledge, providing controlled updates without requiring full retraining. Unlike these *implicit* knowledge expansion methods, which modify the model's internal parameters, retrieval-based approaches *explicitly* access external information dynamically during inference, reducing dependency on static parametric knowledge. The suitability of these methods for different knowledge types and their general applicability are summarized in Table 1. By leveraging these strategies, LLMs can maintain accuracy, contextual awareness, and adaptability to new information.

After placing our work into context (Section 2) and defining knowledge types covered in this paper (Section 3), we provide an overview of different knowledge expansion methods as detailed in Fig-

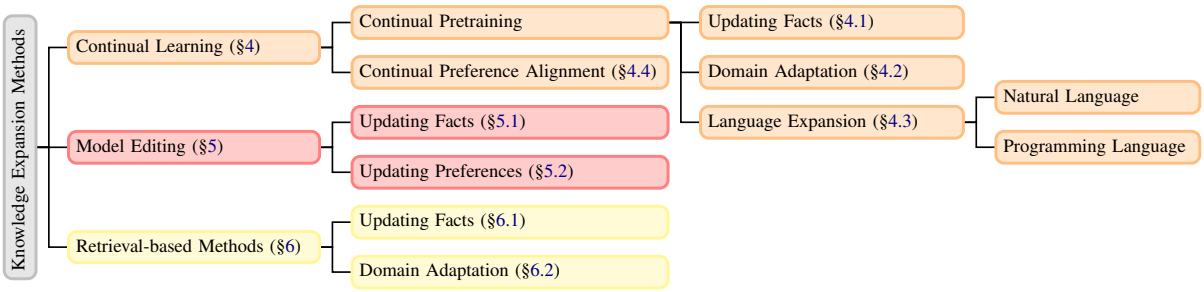

Figure 1: Taxonomy of current methods for expanding LLM knowledge. Due to space constraints, please refer to Appendix A.1 for a comprehensive review of methods and their corresponding citations.

ure 1. This work thus surveys diverse research efforts and may serve as a guide for researchers and practitioners aiming to develop and apply adaptable and robust LLMs. We highlight research opportunities and provide insights into optimizing adaptation techniques for various real-world applications.

2 Related Surveys

The main goal of our work is to provide researchers and practitioners a broad overview of various types of methods to adapt LLMs to diverse types of new knowledge. In this section, we explain how other more specialized surveys relate to our paper.

To the best of our knowledge, there is limited prior work that specifically focuses on continuous knowledge expansion for LLMs. Closest to our work, Zhang et al. (2023c) describe temporal factual knowledge updates, while we take a broader perspective by examining methods for adapting LLMs to unseen domain knowledge, expanding language coverage, and incorporating user preferences. Yao et al. (2023) and Zhang et al. (2024c) provide overviews of knowledge editing methodologies, categorizing approaches of knowledge editing. Similarly, Ke and Liu (2022), Wu et al. (2024b) and Wang et al. (2024b) offer a comprehensive overview of continual learning. In contrast, our survey shifts the focus towards a task-oriented perspective on knowledge expansion, detailing how various types of knowledge — including factual, domain-specific, language, and user preference knowledge — can be seamlessly integrated to ensure LLMs remain relevant and effective.

3 Knowledge Types

Integrating diverse types of knowledge into LLMs is essential to enhance their versatility and effectiveness. Depending on the use case, the type of knowledge that an LLM shall be adapted to, might differ. In this paper, we distinguish four key types of knowledge, which cover a broad range of use cases of researchers and practitioners: **factual, domain, language, and preference knowledge.**

(1) We define **factual knowledge** as general truths or contextualized information about the world that can be expressed in factual statements. We adopt a broad, high-level definition, encompassing finer-grained categorizations, such as commonsense knowledge, cultural knowledge, temporal knowledge, and entity knowledge as subsets of factual knowledge, in contrast to prior works (Cao et al., 2024; Wu et al., 2024b) using more granular classifications. This inclusive perspective enables a comprehensive exploration of knowledge expansion techniques for LLMs, providing flexibility beyond predefined categories and taxonomies.

(2) We define **domain knowledge** as specialized information relevant to specific fields, such as medicine, law, or engineering, enabling LLMs to perform well in targeted applications. Since LLMs typically excel in general-domain tasks but struggle with specialized content, incorporating domain knowledge is crucial for bridging this gap and improving performance in specific fields.

(3) We define **language knowledge** as the ability of an LLM to understand, generate, and reason in specific natural or programming languages.[1] Its integration focuses on adapting models to new languages and enhancing performance in underrepresented ones for broader applicability.

(4) Finally, we define **preference knowledge** as the capability of LLMs to tailor their behavior to align with user-specific needs, preferences, or values. Preference knowledge integration involves

[1]We distinguish language knowledge from linguistic knowledge as defined by Hernandez et al. (2024). Language knowledge refers to the multilingual capabilities of an LLM, whereas linguistic knowledge falls under factual knowledge, encompassing statements about syntax and grammar.

adapting LLM behavior to meet diverse and dynamic user expectations.

In the next sections, we survey knowledge expansion methods and explain for which of these four knowledge types they are suitable.

4 Continual Learning

Continual learning (CL) is a machine learning paradigm that mimics the human ability to continuously learn and accumulate knowledge without forgetting previously acquired information (Chen and Liu, 2018). In the context of knowledge expansion, CL allows LLMs to integrate new corpora and incrementally update the knowledge stored in their parameters. This ensures that LLMs remain adaptable, relevant, and effective as facts, domains, languages, and user preferences evolve over time.

In the era of LLMs, the training of language models usually includes multiple stages: pretraining, instruction tuning, preference alignment, and potentially fine-tuning on a downstream task (Shi et al., 2024a). Depending on the stage, continual learning can be categorized into continual pretraining, continual instruction tuning, continual preference alignment, and continual end-task learning (Ke et al., 2023; Shi et al., 2024a). For knowledge expansion, the focus lies on continual pretraining (CPT) and continual preference alignment (CPA). In contrast, continual instruction tuning and continual end-task learning primarily aim to sequentially fine-tune pretrained LLMs for acquiring new skills and solving new tasks, which fall outside the scope of this survey.

In the following sections, we review existing studies that leverage continual pretraining for updating facts, adapting domains, and expanding languages, and continual alignment for updating user preferences.

4.1 Continual Pretraining for Updating Facts

This line of research focuses on updating a language model's outdated internal factual knowledge by incrementally integrating up-to-date world knowledge (Jang et al., 2022; Ke et al., 2023).

Early studies (Sun et al., 2020; Röttger and Pierrehumbert, 2021; Lazaridou et al., 2021; Dhingra et al., 2022) empirically analyze continual pretraining on temporal data, demonstrating its potential for integrating new factual knowledge. Jin et al. (2022) and Jang et al. (2022) apply traditional continual learning methods to factual knowledge updates in LLMs, evaluating their effectiveness in continual knowledge acquisition. Similarly, Jang et al. (2022) and Kim et al. (2024) classify world knowledge into time-invariant, outdated, and new categories — requiring knowledge retention, removal, and acquisition, respectively — and benchmark existing continual pretraining methods for knowledge updates.

Additionally, Hu et al. (2023) introduce a meta-trained importance-weighting model to adjust per-token loss dynamically, enabling LLMs to rapidly adapt to new knowledge. Yu and Ji (2024) investigate self-information updating in LLMs through continual learning, addressing exposure bias by incorporating fact selection into training losses.

4.2 Continual Pretraining for Domain Adaptation

Continual domain adaptative pretraining (Ke et al., 2022, 2023; Wu et al., 2024b) focuses on incrementally adapting an LLM using a sequence of unlabeled, domain-specific corpora. The objective is to enable the LLM to accumulate knowledge across multiple domains while mitigating catastrophic forgetting (McCloskey and Cohen, 1989) of previously acquired domain knowledge or general language understanding.

Gururangan et al. (2020) introduced the term of domain-adaptive pretraining, demonstrating that a second phase of pretraining on target domains can effectively update an LLM with new domain knowledge. It is important to note that further pretraining can lead to catastrophic forgetting of general concepts by overwriting essential parameters. To mitigate this, recent works utilize *parameter-isolation* methods which allocate different parameter subsets to distinct tasks or domains and keep the majority of parameters frozen (Razdaibiedina et al., 2023; Wang et al., 2024d,e). DEMix-DAPT (Gururangan et al., 2022) replaces every feed-forward network layer in the Transformer model with a domain expert mixture layer, containing one expert per domain. When acquiring new knowledge, only the newly added expert is trained while all others remain fixed. Qin et al. (2022) propose ELLE for efficient lifelong pretraining on various domains. ELLE starts with a randomly initialized LLM and expands the PLM's width and depth to acquire new knowledge more efficiently. Ke et al. (2022) introduce a continual pretraining system which inserts continual learning plugins to the frozen pretrained language models that mitigate catastrophic forget-

ting while effectively learn new domain knowledge. Similarly, Lifelong-MoE (Chen et al., 2023) expands expert capacity progressively, freezing previously trained experts and applying output-level regularization to prevent forgetting.

In a later work, Ke et al. (2023) apply regularization to penalize changes to critical parameters learned from previous data, preventing catastrophic forgetting. Their approach computes the importance of LLM components, such as attention heads and neurons, in preserving general knowledge, applying soft-masking and contrastive loss during continual pretraining to maintain learned knowledge while promoting knowledge transfer.

4.3 Continual Pretraining for Language Expansion

Continual pretraining (CPT) has emerged as a pivotal strategy for adapting LLMs to new languages, or enhancing performance in underrepresented languages without full retraining (Wu et al., 2024b). Below, we discuss two major areas of expansion enabled by CPT: *natural language expansion* and *programming language expansion*.

Natural Language Expansion. Several recent studies have demonstrated the effectiveness of CPT in expanding language coverage. Glot500 (Imani et al., 2023) and EMMA-500 (Ji et al., 2024) enhance multilingual capabilities using CPT and vocabulary extension. Glot500, based on XLM-R (Ruder et al., 2019), and EMMA-500, built on LLaMA 2 (Touvron et al., 2023), expand language support up to 500 languages using extensive multilingual corpora. Similarly, Aya (Üstün et al., 2024) applies continual pretraining to the mT5 model (Xue et al., 2021) using a carefully constructed instruction dataset, achieving improved performance across 101 languages. Furthermore, LLaMAX (Lu et al., 2024) enhances multilingual translation by applying continual pretraining to the LLaMA model family. Supporting over 100 languages, it improves translation quality and promotes language inclusivity.

While covering many languages, many multilingual models exhibit suboptimal performance on medium- to low-resource languages (Ruder et al., 2019; Touvron et al., 2023; Imani et al., 2023). To bridge this performance gap, researchers have focused on expanding training corpora and strategically applying continual pretraining to enhance the multilingual capabilities of LLMs. Alabi et al. (2022), Wang et al. (2023a), Fujii et al. (2024), and Zhang et al. (2024b) show that continual pretraining on one or more specific languages significantly improves performance across related languages. Blevins et al. (2024) extend this approach to the MoE paradigm for better parameter efficiency, while Zheng et al. (2024) investigate scaling laws for continual pretraining by training LLMs of varying sizes under different language distributions and conditions. Additionally, Tran (2020), Minixhofer et al. (2022), Dobler and de Melo (2023), Liu et al. (2024b), and Minixhofer et al. (2024) explore advanced tokenization and word embedding techniques to further improve LLMs' multilingual performance in low-resource settings.

Programming Language Expansion. Going beyond natural languages, continual pretraining has demonstrated significant potential in enhancing the capabilities of LLMs for understanding and generating programming languages.

CERT, proposed by Zan et al. (2022), addresses the challenges of library-oriented code generation using unlabeled code corpora. It employs a two-stage framework to enable LLMs to effectively capture patterns in library-based code snippets. CodeTask-CL (Yadav et al., 2023) offers a benchmark for continual code learning, encompassing a diverse set of tasks such as code generation, summarization, translation, and refinement across multiple programming languages. Furthermore, continual pretrained models specifically for code understanding and programming from natural language prompts emerged with LLMs, such as Code-LLaMA (Grattafiori et al., 2023), Llama Pro (Wu et al., 2024a), CodeGemma (Team et al., 2024) and StarCoder 2 (Lozhkov et al., 2024), consistently outperform general-purpose LLMs of comparable or larger size on code benchmarks.

4.4 Continual Preference Alignment

Preference alignment ensures that large language models generate responses consistent with human values, improving usability, safety, and ethical behavior. While techniques like Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Lambert et al., 2022) align LLMs with static preferences, societal values evolve, requiring continual preference alignment (CPA). It enables LLMs to adapt to emerging preferences while preserving previously learned values, ensuring relevance, inclusivity, and responsiveness to shifting

societal expectations. Despite its importance, CPA remains a relatively underexplored area. Below, we briefly discuss two representative works that highlight the potential of this approach:

Zhang et al. (2023b) propose a non-reinforcement learning approach for continual preference alignment in LLMs. Their method uses function regularization by computing an optimal policy distribution for each task and applying it to regularize future tasks, preventing catastrophic forgetting while adapting to new domains. This provides a single-phase, reinforcement learning-free solution for maintaining alignment across diverse tasks. Zhang et al. (2024a) introduce Continual Proximal Policy Optimization (CPPO), integrating continual learning into the RLHF framework to accommodate evolving human preferences. CPPO employs a sample-wise weighting strategy to balance policy learning and knowledge retention, consolidating high-reward behaviors while mitigating overfitting and noise.

As the demand for responsive and inclusive AI grows, CPA is key to keeping LLMs ethical and aligned with evolving user needs, requiring further research to reach its full potential.

4.5 Applicability and Limitations

Continual learning is a versatile framework for expanding LLM knowledge across facts, domains, languages, and preferences. It excels in large-scale knowledge integration, retaining previously learned knowledge, making it well-suited for tasks like domain adaptation and language expansion (Bu et al., 2021; Jin et al., 2022; Cossu et al., 2024).

However, CL has notable limitations, including a lack of precise control compared to model editing (cf. Section 5) and retrieval-based methods (cf. Section 6), inefficiency due to the computational demands of retraining, and limited applicability in black-box models. These challenges highlight the need for alternative approaches like model editing and retrieval, which offer more targeted and efficient updates.

5 Model Editing

Model editing offers a controllable and efficient solution to update factual knowledge and user preferences in LLMs. Introduced by Zhu et al. (2020), De Cao et al. (2021) and Mitchell et al. (2022a), it aims at modifying the model's predictions for specific inputs without affecting unrelated ones.

Yao et al. (2023) and Zhang et al. (2024c) define four key evaluation metrics for model editing: **(1) reliability**, ensuring the edited model produces the target prediction for the target input; **(2) generalization**, requiring the edited knowledge to apply to all in-scope inputs — inputs that are directly related to the target input, including rephrasings and semantically similar variations; **(3) locality**, preserving original outputs for unrelated out-of-scope inputs; and **(4) portability**, extending the generalization metric by assessing how well updated knowledge transfers to complex rephrasings, reasoning chains, and related facts.

While recent works (Mitchell et al., 2022b; Madaan et al., 2022; Zhong et al., 2023; Zheng et al., 2023) use *model editing* and *knowledge editing* interchangeably for updating factual knowledge, we distinguish between them: model editing is a subset of knowledge editing that modifies model parameters, whereas retrieval-based methods update knowledge dynamically without altering the model's parameters (see Section 6).

5.1 Model Editing for Updating Facts

To address outdated or incorrect information (Lazaridou et al., 2021), model editing research focuses on selectively modifying this knowledge. Below, we highlight key works in this area.

KnowledgeEditor (De Cao et al., 2021) uses a hypernetwork to predict parameter shifts for modifying a fact, trained via constrained optimization for locality. Similarly, MEND (Mitchell et al., 2022a) trains a hypernetwork per LLM layer and decomposes the fine-tuning gradient into a precise one-step parameter update. Given the findings that feed-forward layers in transformers function as key-value memories (Geva et al., 2021), Dai et al. (2022) introduce a knowledge attribution method to identify these neurons and directly modify their values via knowledge surgery.

Recent works employ a locate-and-edit strategy for precise model editing. Using causal tracing, Meng et al. (2022) identify middle-layer feed-forward networks as key to factual predictions and propose ROME, which updates facts by solving a constrained least-squares problem in the MLP weight matrix. MEMIT (Meng et al., 2023) extends ROME to modify thousands of facts simultaneously across critical layers while preserving generalization and locality. BIRD (Ma et al., 2023) introduces bidirectional inverse relationship modeling to mitigate the reverse curse (Berglund et al., 2003)

in model editing. While editing FFN layers has proven effective, PMET (Li et al., 2024d) extends editing to attention heads, achieving improved performance. Wang et al. (2024g) further shift the focus to conceptual knowledge, using ROME and MEMIT to alter concept definitions, finding that concept-level edits are reliable but have limited influence on concrete examples.

5.2 Model Editing for Updating Preferences

Recent works expand model editing beyond factual corrections to aligning LLMs with user preferences, such as ensuring safety, reducing bias, and preserving privacy .

Wang et al. (2024c) use model editing to detoxify LLMs, ensuring safe responses to adversarial inputs and preserving general LLM capabilities, such as fluency, knowledge question answering, and content summarization. Their results show that model editing is promising for detoxification but slightly affects general capabilities. Since LLMs can exhibit social biases (Gallegos et al., 2024), Chen et al. (2024a) propose fine-grained bias mitigation via model editing. Inspired by Meng et al. (2022), they identify key layers responsible for biased knowledge and insert a feed-forward network to adjust outputs with minimal parameter changes, ensuring generalization, locality, and scalability. For privacy protection, Wu et al. (2023) extend Dai et al. (2022)'s work by identifying privacy neurons that store sensitive information. Using gradient attribution, they deactivate these neurons, reducing private data leakage while preserving model performance. Moreover, Mao et al. (2024) apply model editing techniques like MEND to modify personality traits in LLMs, aligning responses to opinion-based questions with target personalities. While effective, this approach can degrade text generation quality.

5.3 Applicability and Limitations

Model editing complements continual learning by allowing fine-grained knowledge updates with lower computational costs. However, research has primarily focused on structured, relational, and instance-level knowledge, with limited exploration of other knowledge types, multilingual generalization, and cross-lingual transfer (Nie et al., 2024; Wei et al., 2025).

Additionally, model editing faces several technical challenges, including limited locality and gradual forgetting in large-scale edits (Bu et al., 2019; Mitchell et al., 2022b; Gupta et al., 2024; Li et al., 2024b), making it more suitable for minor updates. Additionally, it can impact general LLM capabilities (Gu et al., 2024b; Wang et al., 2024f) and downstream performance (Gupta et al., 2024), potentially causing model collapse (Yang et al., 2024b). Addressing these issues will enhance model editing's role alongside continual learning and retrieval, ensuring greater precision in dynamic knowledge adaptation.

6 Retrieval-based Methods

Continual learning and model editing modify a model's parameters to update its internal knowledge, making them implicit knowledge expansion methods (Zhang et al., 2023c). In contrast, retrieval-based methods (Lewis et al., 2020) explicitly integrate external knowledge, allowing models to overwrite outdated or undesired information without parameter modifications. These methods leverage external sources, such as databases, off-the-shelf retriever systems, or the Internet, and thus provide up-to-date or domain-specific knowledge (Zhang et al., 2023c), making them effective for factual updates and domain adaptation.

6.1 Retrieval-based Methods for Updating Facts

Retrieval-based methods enhance LLMs by pairing them with an updatable datastore, ensuring access to current factual information. An early approach, retrieval-augmented generation (RAG) (Lewis et al., 2020), fine-tunes a pre-trained retriever end-to-end with the LLM to improve knowledge retrieval. Similarly, kNN-LM (Khandelwal et al., 2020) interpolates the LLM's output distribution with k-nearest neighbor search results from the datastore, with later works optimizing efficiency (He et al., 2021; Alon et al., 2022) and adapting it for continual learning (Peng et al., 2023b).

For factual knowledge editing, Tandon et al. (2022) store user feedback for post-hoc corrections, while Mitchell et al. (2022b), Madaan et al. (2022), and Dalvi Mishra et al. (2022) retrieve stored edits to guide responses. Chen et al. (2024b) introduce relevance filtering to efficiently handle multiple edits. Retrieval-based in-context learning (Zheng et al., 2023; Ram et al., 2023; Mallen et al., 2023; Yu et al., 2023; Shi et al., 2024b; Bi et al., 2024) enables dynamic factual updates.

For complex reasoning, retrieval supports multi-

hop question answering and iterative prompting: Zhong et al. (2023) propose iterative prompting for multi-hop knowledge editing, while Gu et al. (2024a) use a scope detector to retrieve relevant edits and improve question decomposition via entity extraction and knowledge prompts. Similarly, Shi et al. (2024c) enhance multi-hop question answering by retrieving fact chains from a knowledge graph with mutual information maximization and redundant fact pruning.

In multi-step decision-making, retrieval is combined with Chain-of-Thought (CoT) reasoning (Trivedi et al., 2023; Press et al., 2023). Retrieval also aids post-generation fact-checking and refinement (Gao et al., 2023; Peng et al., 2023a; Song et al., 2024) by revising generated text or prompts based on retrieved facts.

For a more comprehensive review of retrieval-based factual knowledge updates, we refer to Zhang et al. (2023c).

6.2 Retrieval-based Methods for Domain Adaptation

Retrieval-based methods have been widely adopted for various domain-specific tasks, e.g., in science and finance. By integrating retrieved external knowledge, these models enhance their adaptability to specialized domains, improving decision-making, analysis, and information synthesis.

In the biomedical domain, retrieval-based approaches facilitate tasks, such as molecular property identification and drug discovery by integrating structured molecular data and information about biomedical entities like proteins, molecules, and diseases (Wang et al., 2023b; Liu et al., 2023; Wang et al., 2024h; Yang et al., 2024a). For instance, Wang et al. (2023b) and Li et al. (2024a) introduce retrieval-based frameworks that extract relevant molecular data from databases to guide molecule generation. In protein research, retrieval-based approaches enhance protein representation and generation tasks (Ma et al., 2024; Sun et al., 2023). Additionally, Lozano et al. (2023) develop a clinical question-answering system that retrieves relevant biomedical literature to provide more accurate responses in medical contexts.

The finance domain, characterized by its data-driven nature, also benefits from retrieval-based methods (Li et al., 2024f,g). Zhang et al. (2023a) enhance financial sentiment analysis by retrieving real-time financial data from external sources. Furthermore, financial question-answering also benefits from retrieval-based methods, which involves extracting knowledge from professional financial documents. Lin (2024) introduces a PDF parsing method integrated with retrieval-augmented LLMs to retrieve relevant financial insights.

6.3 Applicability and Limitations

Despite their advantages, retrieval-based methods also come with several limitations. A major challenge is their reliance on external knowledge sources, which can introduce inconsistencies or outdated information if not properly curated (Jin et al., 2024; Xu et al., 2024). Their effectiveness also depends on the quality and scope of the retrieval system (Bai et al., 2024; Liu et al., 2024a); poor indexing or noisy retrieval may lead to irrelevant or misleading information. Another key issue is maintaining knowledge consistency across queries. Since retrieval-based methods do not update model parameters, contradictions can arise between retrieved facts and previously generated responses, affecting coherence (Njeh et al., 2024; Zhao et al., 2024; Li et al., 2024c).

Addressing these challenges is essential to improving retrieval-based approaches and ensuring their seamless integration with other LLM adaptation techniques.

7 Challenges, Opportunities, Guidelines

Solving Knowledge Conflicts. An inherent challenge of expanding a model's knowledge is the emergence of knowledge conflicts, which can undermine the consistency and trustworthiness of LLMs (Xu et al., 2024). Studies have identified various types of conflicts following knowledge updates, including (i) temporal misalignment (Luu et al., 2022), where outdated and newly learned facts coexist inconsistently, (ii) model inconsistencies (Huang et al., 2021), where responses to similar queries vary unpredictably, and (iii) hallucinations (Ji et al., 2023), where the model generates fabricated or contradictory information. While some efforts have been made to address these issues (Zhang and Choi, 2023; Mallen et al., 2023; Zhou et al., 2023; Xie et al., 2024), they remain an open challenge that requires further research and more robust solutions.

Minimizing Side Effects. Continual learning and model editing, both of which involve modifying model parameters, inevitably introduce side effects. A major challenge in continual learning

is catastrophic forgetting (McCloskey and Cohen, 1989), where newly acquired knowledge overwrites previously learned information. In LLMs, the multi-stage nature of training exacerbates this issue, leading to cross-stage forgetting (Wu et al., 2024b), where knowledge acquired in earlier stages is lost as new training phases are introduced. For model editing, recent studies have shown that large-scale edits, particularly mass edits, can significantly degrade the model's general capabilities, such as its language modeling performance (Wang et al., 2024f) or accuracy on general NLP tasks (Li et al., 2024e,b; Wang et al., 2024a). Effectively addressing these challenges is crucial for maximizing the potential of these methods for large-scale knowledge expansion while maintaining model stability and overall performance.

Comprehensive Benchmarks. Although this paper explores the properties, strengths, and weaknesses of various methods for knowledge expansion, the discussion remains largely theoretical due to the lack of a comprehensive benchmark datasets for a uniform evaluation and a proper comparison. Existing works, such as Jang et al. (2022), Liska et al. (2022), and Kim et al. (2024), provide factual knowledge-based datasets and evaluate continual pretraining and/or retrieval-based methods. However, their experiments are limited in scale and fail to offer a comprehensive assessment. Developing benchmarks that encompass a variety of knowledge types and enable the evaluation of all methods would provide a more holistic and systematic understanding of their relative effectiveness.

General Guideline. Selecting the appropriate method for knowledge expansion in LLM depends on the application context and type of knowledge that needs to be updated.

(i) For **factual knowledge**, model editing is ideal for precise, targeted updates, such as correcting specific facts, due to its efficiency and high level of control. Retrieval-based methods are effective for integrating dynamic or frequently changing facts, as they allow updates without modifying the model's parameters, making them suitable for black-box applications. For large-scale factual updates, continual learning is preferred as it enables the incremental integration of new knowledge while preserving previously learned information.

(ii) For **domain knowledge**, both continual learning and retrieval-based methods are applicable. Continual learning excels in large-scale adaptation, using domain-specific corpora to ensure the model retains general knowledge while adapting to specialized contexts. Retrieval-based methods complement this by dynamically providing domain-specific information without requiring model modifications, making them valuable in scenarios where static updates are impractical.

(iii) For **language knowledge**, continual learning is the only method capable of supporting large-scale language expansion. It facilitates the integration of multilingual corpora and provides the foundational updates necessary for underrepresented or low-resource languages.

(iv) For **preference updates**, such as aligning models with evolving user values or ethical norms, continual alignment is typically achieved by combining continual learning techniques with preference optimization methods, such as reinforcement learning from human feedback. These approaches enable models to dynamically adapt to changing preferences while retaining alignment with previously learned values.

Summary. Continual learning is indispensable for large-scale updates like domain adaptation and language expansion, where foundational and incremental updates are required. **Model editing** excels at precise factual updates, while **retrieval-based methods** offer dynamic access to factual and domain knowledge without altering the model. A well-informed selection or combination of these methods ensures efficient and effective knowledge expansion tailored to specific use cases.

8 Conclusions

Adapting large language models to evolving knowledge is essential for maintaining their relevance and effectiveness. This survey explores three key adaptation methods — continual learning for large-scale updates, model editing for precise modifications, and retrieval-based approaches for external knowledge access without altering model parameters. We examine how these methods support updates across factual, domain-specific, language, and user preference knowledge while addressing challenges like scalability, controllability, and efficiency. By consolidating research and presenting a structured taxonomy, this survey provides insights into current strategies and future directions, promoting the development of more adaptable and efficient large language models.

Limitations

This survey provides a comprehensive overview of knowledge expansion techniques for LLMs. However, due to page constraints, we had to limit its scope and prioritize certain aspects:

First, the paper only provides a high-level overview of each method rather than an in-depth analysis. This can limit the understanding of the nuances and specific applications of each technique, as well as implementation details.

Second, our work is a literature review of adaptation methods rather than an empirical study evaluating their actual performance. While we analyze existing strategies, we do not benchmark or experimentally compare their effectiveness, leaving room for future studies to assess their practical impact under real-world conditions.

Third, we focus solely on text-based models and do not cover vision-language models, which integrate multi-modal learning for textual and visual understanding. While the methods covered in this survey could be used to adapt the language encoders of such models in theory, extending these adaptation methods to vision-language models remains an open research direction.

Finally, this survey reflecting the current state of research might become outdated as new research is published, as the field of LLMs is rapidly evolving and new methods for knowledge expansion are continuously being developed.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 468–485. PMLR.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2003. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *arXiv preprint arXiv:2405.11613*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837, Miami, Florida, USA. Association for Computational Linguistics.

Xingyuan Bu, Junran Peng, Junjie Yan, Tieniu Tan, and Zhaoxiang Zhang. 2021. GAIA: A Transfer Learning System of Object Detection that Fits Your Needs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 274–283. IEEE Computer Society.

Xingyuan Bu, Yuwei Wu, Zhi Gao, and Yunde Jia. 2019. Deep convolutional network with locality and sparsity constraints for texture classification. *Pattern Recognition*, 91:34–46.

Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2024. The life cycle of knowledge in big language models: A survey. *Machine Intelligence Research*, 21(2):217–238.

Ruizhe Chen, Yichen Li, Zikai Xiao, and Zuozhu Liu. 2024a. Large language model bias mitigation from the perspective of knowledge editing. *arXiv preprint arXiv:2405.09341*.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023. Lifelong language pretraining with distribution-specialized experts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5383–5395. PMLR.

Yingfa Chen, Zhengyan Zhang, Xu Han, Chaojun Xiao, Zhiyuan Liu, Chen Chen, Kuai Li, Tao Yang, and Maosong Sun. 2024b. Robust and scalable model editing for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14157–14172, Torino, Italia. ELRA and ICCL.

Zhiyuan Chen and Bing Liu. 2018. Continual learning and catastrophic forgetting. In *Lifelong Machine Learning*, pages 55–75. Springer.

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. 2024. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Bhavana Dalvi Mishra, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9465–9480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenhan Xiong Grattafiori, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. 2024a. PokeMQA: Programmable knowledge editing for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8069–8083, Bangkok, Thailand. Association for Computational Linguistics.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024b. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16801–16819, Miami, Florida, USA. Association for Computational Linguistics.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15202–15232, Bangkok, Thailand. Association for Computational Linguistics.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. 2022. DEMix layers: Disentangling domains for modular language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5557–5576, Seattle, United States. Association for Computational Linguistics.

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

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5703–5714, Online and Punta Cana,

Dominican Republic. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Nathan Hu, Eric Mitchell, Christopher Manning, and Chelsea Finn. 2023. Meta-learning online adaptation of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4418–4432, Singapore. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations*.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Jiajie Jin, Yutao Zhu, Yujia Zhou, and Zhicheng Dou. 2024. BIDER: Bridging knowledge inconsistency for efficient retrieval-augmented LLMs via key supporting evidence. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 750–761, Bangkok, Thailand. Association for Computational Linguistics.

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 *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780, Seattle, United States. Association for Computational Linguistics.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10216, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. 2024. Carpe diem: On the evaluation of world knowledge in lifelong language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5401–5415, Mexico City, Mexico. Association for Computational Linguistics.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*. Https://huggingface.co/blog/rlhf.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Curran Associates Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Curran Associates Inc.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024a. Empowering Molecule Discovery for Molecule-Caption Translation With Large Language Models: A ChatGPT Perspective . *IEEE Transactions on Knowledge & Data Engineering*, 36(11):6071–6083.

Qi Li, Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Xinglin Pan, and Xiaowen Chu. 2024b. Should we really edit language models? on the evaluation of edited language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. 2024c. GraphReader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12758–12786, Miami, Florida, USA. Association for Computational Linguistics.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024d. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18564–18572.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2024e. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations*.

Zichao Li, Bingyang Wang, and Ying Chen. 2024f. Incorporating economic indicators and market sentiment effect into us treasury bond yield prediction with machine learning. *Journal of Infrastructure, Policy and Development*, 8(9):7671.

Zichao Li, Bingyang Wang, and Ying Chen. 2024g. Knowledge graph embedding and few-shot relational learning methods for digital assets in usa. *Journal of Industrial Engineering and Applied Science*, 2(5):10–18.

Demiao Lin. 2024. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition. *arXiv preprint arXiv:2401.12599*.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Dong Liu, Roger Waleffe, Meng Jiang, and Shivaram Venkataraman. 2024a. Graphsnapshot: Graph machine learning acceleration with fast storage and retrieval. *arXiv preprint arXiv:2406.17918*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multimodal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024b. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 8–23. World Scientific.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Chang Ma, Haiteng Zhao, Lin Zheng, Jiayi Xin, Qintong Li, Lijun Wu, Zhihong Deng, Yang Young Lu, Qi Liu, Sheng Wang, and Lingpeng Kong. 2024. Retrieved sequence augmentation for protein representation learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1738–1767, Miami, Florida, USA. Association for Computational Linguistics.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. 2023. Untying the reversal curse via bidirectional language model editing. *arXiv preprint arXiv:2310.10322*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2833–2861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part II*, page 241–254. Springer-Verlag.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

Ercong Nie, Bo Shao, Zifeng Ding, Mingyang Wang, Helmut Schmid, and Hinrich Schütze. 2024. Bmike-53: Investigating cross-lingual knowledge editing with in-context learning. *arXiv preprint arXiv:2406.17764*.

Chaima Njeh, Haïfa Nakouri, and Fehmi Jaafar. 2024. Enhancing RAG-retrieval to improve LLMs robustness and resilience to hallucinations. In *Hybrid Artificial Intelligent Systems*, pages 201–213. Springer Nature Switzerland.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2024. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023a. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Guangyue Peng, Tao Ge, Si-Qing Chen, Furu Wei, and Houfeng Wang. 2023b. Semiparametric language models are scalable continual learners. *arXiv preprint arXiv:2303.01421*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

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 *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2789–2810, Dublin, Ireland. Association for Computational Linguistics.

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

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024a. Continual learning of large language models: A comprehensive survey.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024b. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. 2024c. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2056–2066. Association for Computing Machinery.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Xiaoshuai Song, Zhengyang Wang, Keqing He, Guanting Dong, Yutao Mou, Jinxu Zhao, and Weiran Xu. 2024. Knowledge editing on black-box large language models. *arXiv preprint arXiv:2402.08631*.

Fang Sun, Zhihao Zhan, Hongyu Guo, Ming Zhang, and Jian Tang. 2023. Graphvf: Controllable protein-specific 3d molecule generation with variational flow. *arXiv preprint arXiv:2304.12825*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 339–352, Seattle, United States. Association for Computational Linguistics.

CodeGemma Team, Heri Zhao, Jeffrey Hui, Joshua Howland, Nam Nguyen, Siqi Zuo, Andrea Hu, Christopher A Choquette-Choo, Jingyue Shen, Joe Kelley, et al. 2024. Codegemma: Open code models based on gemma. *arXiv preprint arXiv:2406.11409*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Jianchen Wang, Zhouhong Gu, Zhuozhi Xiong, Hongwei Feng, and Yanghua Xiao. 2024a. The missing piece in model editing: A deep dive into the hidden damage brought by model editing. *arXiv preprint arXiv:2403.07825*.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024b. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024c. Detoxifying large language models via knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3093–3118, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schuetze. 2024d. Learn it or leave it: Module composition and pruning for continual learning. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 163–176, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schuetze. 2024e. Rehearsal-free modular and compositional continual learning for language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 469–480, Mexico City, Mexico. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023a. NLNDE at SemEval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 488–497, Toronto, Canada. Association for Computational Linguistics.

Mingyang Wang, Lukas Lange, Heike Adel, Jannik Strötgen, and Hinrich Schuetze. 2024f. Better call SAUL: Fluent and consistent language model editing with generation regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7990–8000, Miami, Florida, USA. Association for Computational Linguistics.

Xiaohan Wang, Shengyu Mao, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. 2024g. Editing conceptual knowledge for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 706–724, Miami, Florida, USA. Association for Computational Linguistics.

Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. 2023b. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations*.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N. Ioannidis, Huzefa Rangwala, and RISHITA ANUBHAI. 2024h. Biobridge: Bridging biomedical foundation models via knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. MLaKE: Multilingual knowledge editing benchmark for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4457–4473, Abu Dhabi, UAE. Association for Computational Linguistics.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ying Shan, and Ping Luo. 2024a. LLaMA pro: Progressive LLaMA with block expansion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6518–6537, Bangkok, Thailand. Association for Computational Linguistics.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024b. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Prateek Yadav, Qing Sun, Hantian Ding, Xiaopeng Li, Dejiao Zhang, Ming Tan, Parminder Bhatia, Xiaofei Ma, Ramesh Nallapati, Murali Krishna Ramanathan, Mohit Bansal, and Bing Xiang. 2023. Exploring continual learning for code generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 782–792, Toronto, Canada. Association for Computational Linguistics.

Ling Yang, Zhilin Huang, Xiangxin Zhou, Minkai Xu, Wentao Zhang, Yu Wang, Xiawu Zheng, Wenming Yang, Ron O. Dror, Shenda Hong, and Bin CUI. 2024a. Prompt-based 3d molecular diffusion models for structure-based drug design.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Du Su, Dawei Yin, and Huawei Shen. 2024b. The fall of ROME: Understanding the collapse of LLMs in model editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4079–4087, Miami, Florida, USA. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*.

Pengfei Yu and Heng Ji. 2024. Information association for language model updating by mitigating LM-logical discrepancy. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 117–129, Miami, FL, USA. Association for Computational Linguistics.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2436, Toronto, Canada. Association for Computational Linguistics.

Daoguang Zan, Bei Chen, Dejian Yang, Zeqi Lin, Minsu Kim, Bei Guan, Yongji Wang, Weizhu Chen, and Jian-Guang Lou. 2022. CERT: Continual pre-training on sketches for library-oriented code generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2369–2375. International Joint Conferences on Artificial Intelligence Organization.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, pages 349–356. Association for Computing Machinery.

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

Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. 2024a. Cppo: Continual learning for reinforcement learning with human feedback. In *The Twelfth International Conference on Learning Representations*.

Miaoran Zhang, Mingyang Wang, Jesujoba Alabi, and Dietrich Klakow. 2024b. AAdaM at SemEval-2024 task 1: Augmentation and adaptation for multilingual semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 800–810, Mexico City, Mexico. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226, Singapore. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024c. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023c. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4).

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7725–7738, Miami, Florida, USA. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Comprehensive Taxonomy of Methods

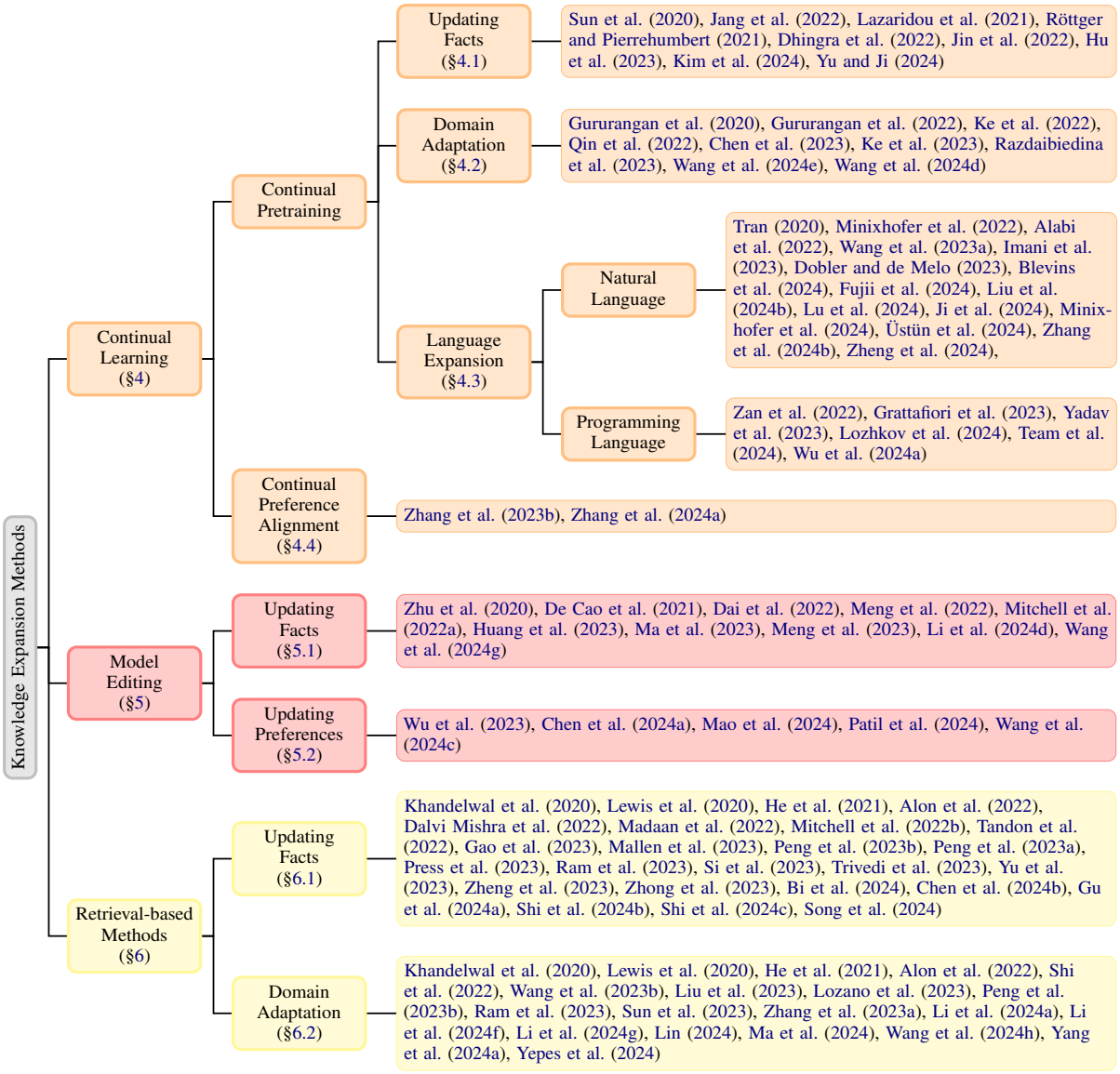

Figure 2: Taxonomy of methods for expanding LLM knowledge.