Advancements in Representation Learning and Personalization of Large Language Models (2023-2025)

Executive Summary

This report provides a concise overview of the significant advancements in Large Language Models (LLMs) concerning representation learning and personalization from 2023 to 2025. It highlights the rapid expansion of LLM research, marked by an exponential increase in publications and a notable shift in institutional leadership from industry to academia. The analysis delves into critical improvements in representation learning, including the evolution from large to smaller, more efficient models, data-centric approaches for enhancing representation quality, and sophisticated memory mechanisms that enable dynamic and long-term understanding. Furthermore, the report explores the imperative for personalization, detailing a comprehensive taxonomy of techniques spanning input-level prompting, model-level adaptation (including collaborative Parameter-Efficient Fine-Tuning), and objective-level alignment through representation editing. Practical applications across conversational AI, recommendation systems, e-commerce, and education are examined, showcasing how personalized LLMs are enhancing user experience and domain-specific utility. The report concludes by outlining persistent technical challenges such as efficiency, data sparsity, and complex data integration, alongside crucial ethical considerations related to privacy, bias, and trustworthiness. It projects a future where LLMs are not only powerful but also deeply personal, adaptive, and responsibly integrated into human lives through continuous lifelong learning.

1. Introduction: The Evolving Landscape of Large Language Models

1.1 The Transformative Impact of LLMs

Large Language Models (LLMs) have ushered in a new era of artificial intelligence, demonstrating remarkable capabilities across a broad spectrum of natural language processing (NLP) tasks. These include sophisticated open-domain dialogue, precise question answering, advanced content generation, and efficient task assistance.[1] Their proficiency in generating, comprehending, and adapting to human language has propelled significant advancements in artificial general intelligence (AGI), fundamentally reshaping human-computer interactions and professional workflows.[3]

The rapid development and deployment of LLMs, such as GPT-4, Mixtral 8x22B, PaLM-340B, and LLaMA-405B, have led to their pervasive presence in daily life. By March 2024, for instance, ChatGPT alone had garnered approximately 180 million users.[1] This widespread adoption has coincided with an exponential surge in research activities, fundamentally altering research priorities across diverse computer science conferences and fields.[3] The sheer volume of new studies indicates a rapid maturation and mainstreaming of LLM research, transforming it from a specialized area into a foundational pillar across various computer science subfields. This implies a fundamental shift where LLMs are increasingly viewed as core enablers rather than mere applications, leading to accelerated innovation but also to challenges in maintaining research quality and reproducibility amidst the high volume of publications.

1.2 Defining Representation Learning and Personalization in the LLM Era

At the core of LLM capabilities lies **representation learning**. This refers to the intricate process by which LLMs internally process, encode, and understand information. It involves the acquisition of general language representations during an extensive pre-training phase, followed by the transfer of this learned knowledge to enhance performance on specific NLP tasks through fine-tuning.[1] The effectiveness and efficiency of these learned representations are paramount to the overall performance of any LLM system.[7]

Complementing representation learning is **personalization**, a critical advancement that tailors LLM outputs and behaviors to individual user preferences, historical interactions, and specific contexts.[8] While LLMs excel at general knowledge tasks,

they often encounter limitations in user-specific personalization, struggling to fully grasp individual emotions, writing styles, and unique preferences.[11] Personalized LLMs (PLLMs) directly address these challenges by leveraging granular user data, including profiles, historical dialogues, content, and interactions, to deliver responses that are contextually relevant and uniquely tailored to each user. This capability significantly enhances user satisfaction and broadens the applicability of LLMs across various domains.[11]

1.3 Report Scope: Navigating Recent Research (2023-2025)

This report synthesizes cutting-edge research from leading academic conferences and pre-print archives, including ACL, EMNLP, NeurIPS, ICML, ICLR, ArXiv, SIGIR, AAAI, and KDD.[15] The focus is on publications from 2023 to 2025, providing a comprehensive overview of the synergistic advancements in LLM representation learning and personalization during this pivotal period.

2. Macro Trends in LLM Research (2023-2025)

2.1 Exponential Growth and Shifting Research Priorities

The landscape of LLM research has undergone a dramatic transformation, characterized by an exponential increase in publications. In 2019, only 503 LLM-related papers were recorded, signifying the nascent stage of this research area. This number doubled by 2020 and continued to climb in 2021. The most substantial surge occurred in 2024, with 7,109 LLM-related papers published, marking an increase of 3,255 papers compared to 2023.[3] This dramatic growth underscores the central and rapidly advancing role of LLMs in computer science research. The sheer volume of publications indicates that LLM research has moved from a specialized sub-discipline to a foundational area influencing a wide array of computer science fields, including AI, systems, and various interdisciplinary domains.[3] This profound shift in research priorities suggests that LLMs are now considered fundamental enablers across the

computing landscape.

Furthermore, LLM research is actively driving significant topic shifts within major conferences. Natural Language Processing (NLP) conferences, such as ACL, EMNLP, and NAACL, consistently prioritize research on LLM adaptation, evaluation, and core development, with a particular emphasis on embedding techniques.[3] Concurrently, machine learning conferences, including ICLR, ICML, and NeurIPS, focus heavily on architectural and efficiency improvements for LLMs. This includes techniques like compression, sparsity, quantization, and Parameter-Efficient Fine-Tuning (PEFT).[3] This division of labor creates a complementary research ecosystem where NLP researchers refine how LLMs understand and represent language, which is crucial for task performance, while ML researchers optimize the underlying models and their efficiency, vital for scalability and practical deployment. This complementary specialization fosters a holistic advancement of LLMs, ensuring that both theoretical foundations (representation quality) and practical considerations (efficiency, scalability) are rigorously addressed. Future breakthroughs may well emerge from the intersection of these two areas, such as developing more efficient methods for learning richer, task-specific representations.

2.2 Global Contributions: Institutions and Nations Driving Innovation

The leadership in LLM research has seen a notable evolution, with a discernible shift from industry giants to academic powerhouses. From 2019 through the early 2020s, major technology companies like Google, Microsoft, Meta, and Amazon predominantly led LLM research publications. Google, for instance, consistently held a top-tier position from 2019 to 2023, driven by its extensive computational resources, proprietary data, and substantial investments in LLM development.[3] However, by 2024, academic institutions such as Tsinghua University (THU), Nanyang Technological University (NTU), Stanford, the University of Washington (UW), and the Hong Kong University of Science and Technology (HKUST) have steadily gained prominence, challenging and even surpassing their industry counterparts.[3] This increased academic leadership suggests a democratization of LLM development, likely facilitated by the growing availability and maturity of open-source LLMs like LLaMA, BERT, Gemma, and Phi.[1] The reduced prohibitive computational and data costs associated with these open-source models enable academic researchers with limited resources to engage in cutting-edge LLM research, fostering a more diverse and collaborative research environment. This trend can lead to more varied research

directions, faster innovation cycles, and potentially more accessible and ethically sound LLM technologies. It also implies that the "black box" nature of LLMs [12] may be more rigorously addressed through academic scrutiny and open-source contributions, leading to more transparent and trustworthy AI systems, and a shift in focus from merely scaling up models to optimizing smaller, more efficient models for specific applications.[1]

The global landscape of LLM research is dominated by the United States and China, which consistently hold the first and second positions in research output, respectively, reflecting their sustained global influence. The United Kingdom consistently ranks third, while Hong Kong has shown significant advancement, rising to fourth position by 2024. Other notable contributors include South Korea, Singapore, Germany, India, and Canada, maintaining stable positions within the top ten.[3] While these leading nations share core LLM research priorities such as efficient LLM adaptation, LLM reasoning, and mitigating hallucinations, they also exhibit distinct thematic specializations. For instance, the United States emphasizes LLM application, prompting, and "LLM for Embedding." China focuses on vision-language models, and the United Kingdom prioritizes "LLM for Embedding" most highly.[3] These specializations are hypothesized to stem from unique national interests in specific application domains, such as robotics in the US, multimodal systems in China, and neural machine translation and medical applications in the UK.

The following tables provide a summary of the top contributing institutions and leading nations, highlighting the evolving landscape of LLM research.

Table 1: Prominent Contributing Institutions to LLM Research (2024)

Institution Type	Leading Entities (2024)	Key Research Focus (as observed)
Academic	Tsinghua University (THU)	General LLM research leadership
	Nanyang Technological University (NTU)	General LLM research leadership
	Stanford University	General LLM research leadership, Socially Aware NLP, Human-AI Interaction [19]

	University of Washington (UW)	General LLM research leadership
	Hong Kong University of Science and Technology (HKUST)	General LLM research leadership
	Zhejiang University (ZJU)	General LLM research leadership
	University of Wisconsin-Madison	Data-driven systems, foundation models, automated ML, data-centric AI, representation tradeoffs, personalization via representation editing [21]
	Oxford Internet Institute (OII)	Responsible personalization, human feedback, aligning LLMs to individuals [23]
	UMass Amherst (CIIR)	Neural Information Retrieval, Conversational Search, Retrieval-Enhanced Machine Learning [24]
	The Chinese University of Hong Kong	Memory in AI, dialogue systems, multilingual confidence estimation [26]
	University of Edinburgh	Memory in AI, dialogue systems [26]
Industry	Google	Early dominance, continued significant contributions [3]
	Microsoft	Early dominance, continued significant contributions [3]
	Meta	Early dominance, continued significant contributions [3]
	Amazon	Early dominance, continued significant contributions [3]

	Adobe Research	Graph representation learning, multilingual embeddings, long document understanding, personalization via heterogeneous feedback [27]
	Apple Machine Learning Research	LLM personalization, remembering user conversations, parameter-efficient settings [29]
	Snorkel AI	Data-first approach to AI, foundation models, enterprise alignment, efficient LLM training [22]

Table 2: Leading Nations in LLM Research and Thematic Specializations

Nation	Overall Output Rank (Consistent)	Key Thematic Specializations (Examples)
United States	1st	LLM application, prompting, LLM for embedding (representation learning), robotics [3]
China	2nd	Vision-language models, multimodal systems [3]
United Kingdom	3rd	High prioritization of "LLM for Embedding" (representation learning), neural machine translation, medical applications [3]
Hong Kong	4th (by 2024)	General LLM research leadership [3]
South Korea	Stable Top 10	General LLM research contributions [3]
Singapore	Stable Top 10	General LLM research

		contributions [3]
Germany	Stable Top 10	General LLM research contributions [3]
India	Stable Top 10	General LLM research contributions [3]
Canada	Stable Top 10	General LLM research contributions [3]

3. Advancements in Representation Learning for LLMs

3.1 Foundational Paradigms: Pre-training and Fine-tuning

The bedrock of modern LLM capabilities remains the "pre-train and fine-tune" paradigm. This approach involves a two-stage process: first, learning broad, general language representations through extensive pre-training on vast and diverse datasets; and second, transferring this acquired knowledge to enhance performance on specific NLP tasks through targeted fine-tuning.[1] This methodology has consistently yielded exceptional performance across a wide array of tasks, including sophisticated language generation, nuanced language understanding, and highly specialized domain-specific applications in fields such as coding, medicine, and law.[1] LLMs are increasingly recognized as versatile, general-purpose learners due to their inherent flexibility to adapt to new tasks, often requiring only a fraction of the training data that was historically necessary for comparable performance.[6]

3.2 Enhancing Efficiency: From Large to Small Language Models

The escalating computational costs and significant energy consumption associated with scaling up LLM sizes, exemplified by models like GPT-4 and LLaMA-405B, have spurred a critical shift in research focus towards smaller language models (SLMs).[1]

This transition is driven by the recognition that SLMs, such as Phi-3.8B and Gemma-2B, can achieve performance comparable to their larger counterparts while operating with substantially fewer parameters.[1] This growing emphasis on SLMs and efficiency techniques represents a crucial evolution in LLM representation learning, moving beyond sheer scale to prioritize practical deployability and resource optimization. The drive for efficiency directly addresses the substantial computational resources required for both training and inference, which often render very large models impractical for academic researchers and businesses with limited resources.[1] This shift makes LLMs more accessible for real-time applications, edge devices, and broader adoption beyond the confines of large technology corporations, indicating that the utility of LLMs is now as important as their raw scale.

Research indicates that smaller models, particularly BERT-base, maintain high popularity in practical settings, suggesting that their utility is often underestimated.[1] While LLMs are celebrated for their broad generalizability, studies show that fine-tuning SLMs on domain-specific datasets can, in certain cases, surpass the performance of general LLMs for highly specialized tasks.[1] This highlights the value of targeted efficiency.

To further enhance efficiency, machine learning conferences are dedicating significant attention to architectural improvements of LLMs. A major emphasis is placed on optimizing Transformer architectures and LLM efficiency through various techniques, including compression, sparsity, quantization, and Parameter-Efficient Fine-Tuning (PEFT).[5] PEFT methods, such as LoRA, are particularly crucial for efficient personalization, as they enable the adaptation of model parameters without the need for extensive retraining of the entire model.[9] Recent work in this area includes explorations into low-bit quantization for LLM compression (featured at ICML 2024 and NeurIPS 2023), memory-efficient fine-tuning via sub-4-bit integer quantization (NeurIPS 2023), and activation-aware weight quantization (recognized with a Best Paper award at MLSys 2024).[32] This trend opens new research avenues in optimizing model architectures for smaller footprints, developing novel PEFT methods (e.g., Per-Pcs for collaborative PEFT [9]), and exploring hybrid approaches where LLMs can collaborate with SLMs to strike a balance between power and efficiency.[1] It also suggests a future where personalized LLMs can be deployed more widely, even on resource-constrained devices, by leveraging these efficiency gains.

3.3 Data-Centric Approaches to Representation Quality

The sophisticated reasoning capabilities observed in LLMs are largely attributed to their pre-training on extensive and diverse datasets, such as C4 and Pile, typically sourced from web scrapes, books, and scientific literature.[1] However, recent research is challenging the notion that sheer data quantity is the sole determinant of performance. A growing body of work supports the idea that "less is more," advocating for advanced data selection or pruning techniques. These methods aim to curate high-quality subsets from large datasets, thereby enhancing model performance and the quality of learned representations.[1] This indicates a strategic shift towards prioritizing data quality over mere volume in the pursuit of superior representations.

Beyond curation, LLMs themselves are being leveraged to generate synthetic data for machine learning tasks, extending beyond traditional language processing. This innovative application enables the creation of higher-quality datasets that more accurately reflect the true complexity of target tasks and distributions.[6] This ability to self-generate and refine training data holds significant promise for continually improving the robustness and specificity of LLM representations.

3.4 Memory Mechanisms and Dynamic Representations

Memory is a fundamental component of advanced AI systems, serving as the underpinning for LLM-based agents and enabling them to sustain coherent and long-term interactions.[2] While earlier surveys on LLM memory often focused on applications, recent work emphasizes the atomic operations that govern memory dynamics.[33] The explicit categorization and operationalization of memory mechanisms in LLMs, particularly the distinction between parametric and contextual memory, is crucial for developing robust and dynamically adaptive personalized LLMs.

Memory representations are broadly categorized into two main forms:

- **Parametric Memory:** This refers to the knowledge implicitly stored within a model's internal parameters. Acquired during pre-training or subsequent post-training, this memory is embedded in the model's weights and accessed through feedforward computation during inference.[34]
- **Contextual Memory:** This encompasses explicit external information, which can be either structured (e.g., knowledge graphs, relational tables, ontologies) or

unstructured (e.g., multi-turn dialogue history, observations from external environments). Contextual memory is modality-general, capable of storing and retrieving information across heterogeneous inputs such as text, images, audio, and video.[34]

Accompanying these memory types are six fundamental memory operations:

- **Consolidation:** The process of integrating new knowledge into persistent memories.[33]
- **Updating:** Modifying existing memory in response to new data or evolving information.[33]
- **Indexing:** Efficiently organizing memory content to facilitate rapid and accurate retrieval.[33]
- **Forgetting:** The deliberate or implicit process of discarding or de-emphasizing old, irrelevant, or noisy information to prevent memory overload and maintain efficiency.[2]
- **Retrieval:** The act of accessing relevant memory content when needed for generating responses or making decisions.[33]
- **Compression:** Reducing the size of memory while preserving essential information, crucial for efficient storage and reasoning, especially in resource-constrained environments.[33]

The structured view of memory provided by this taxonomy directly addresses the challenge of LLMs struggling to persistently remember and incorporate user-specific preferences in a long-term context.[2] Memory storage for LLMs is an increasingly active area of research, particularly for enabling personalization across extended conversations. LLMs are currently limited in their ability to persistently remember and incorporate user-specific preferences in a long-term context.[2] Solutions like HippoRAG 2 are being developed to endow LLMs with more human-like memory capabilities, aiming to overcome limitations that arise when the volume of information grows and tasks become more complex.[7] Retrieval-Augmented Generation (RAG) has emerged as a scalable and practical alternative for continual learning, allowing LLMs to retrieve relevant external information at inference time rather than requiring internal model modification.[7] Improvements in encoder models, particularly those leveraging LLM backbones, significantly enhance RAG systems by generating high-quality embeddings that better capture semantic relationships, thereby improving retrieval quality for LLM generation.[7] Personalized memory construction involves designing mechanisms for retaining and updating memory for efficient retrieval, distinguishing between non-parametric (token-based database) and parametric (learnable space projection) memory.[11] Effective memory management, especially retrieval and

updating of personalized contextual memory, is paramount for personalized LLMs. It enables lifelong learning, where models continually adapt to evolving user behaviors without catastrophic forgetting, and supports complex agents that need to retain information across multiple sessions.[36] This understanding is foundational for building truly adaptive and personalized AI.

3.5 Multimodal Representation Learning

The capabilities of LLMs are expanding beyond text to encompass multimodal data, including audio, images, and video.[4] This evolution is driven by research into multimodal multitask learning using unified transformers [37] and efforts to scale visual and vision-language representation learning.[37] The increasing prominence of multimodal LLMs is reflected in the rising number of LLM-related papers presented at computer vision conferences such as CVPR and ECCV in 2024.[3]

The expansion of LLMs into multimodal representation learning signifies a profound move towards more human-like, comprehensive understanding. Human interaction is inherently multimodal, and by integrating text, audio, and visual data, LLMs can form richer, more complete representations of the world and user intent. This transcends purely linguistic understanding, allowing for the incorporation of contextual cues from other modalities, leading to more nuanced and accurate interpretations of user input and preferences. For personalization, this means the ability to tailor responses not just to textual style but also to visual cues (e.g., inferring user fashion preferences from images) or auditory patterns (e.g., recognizing tone of voice). This holistic understanding directly enhances personalization by enabling LLMs to build a more comprehensive "profile" of a user based on all their interactions. This also opens up new applications, such as personalized content generation across modalities, multimodal recommendation systems, and more natural, empathetic conversational AI.[12] However, it also introduces challenges in aligning representations across diverse modalities and ensuring coherence in multimodal outputs.

4. Personalization of Large Language Models: Techniques and Frameworks

4.1 The Imperative for Tailored LLM Interactions

The ability to personalize Large Language Models has rapidly gained importance, leading to a wide array of applications that aim to tailor interactions, content, and recommendations to individual user preferences.[8] Despite their general proficiency, LLMs often fall short in user-specific personalization, struggling to accurately capture individual emotions, writing styles, and nuanced preferences.[11] Personalized Large Language Models (PLLMs) are specifically designed to overcome these limitations by leveraging individual user data, such as profiles, historical dialogues, content, and interactions. This enables PLLMs to deliver responses that are contextually relevant and uniquely tailored to each user's specific needs, significantly enhancing user satisfaction and the overall utility of the models.[11] The increasing demand for personalized LLM-based conversational agents is evident, with users actively seeking customization features to align the system with their specific usage goals and preferences.[12]

4.2 A Comprehensive Taxonomy of Personalization Techniques

Recent surveys have systematically unified the diverse literature on personalized LLMs, proposing comprehensive taxonomies that categorize personalization techniques, datasets, evaluation methods, and applications, as well as the granularity of personalization.[8] One such taxonomy broadly classifies PLLM methods into three major technical levels [11]: input-level, model-level, and objective-level personalization.

4.2.1 Input-Level Personalization (Prompting-based)

This category focuses on managing user-specific data externally and dynamically injecting it into the LLM, primarily through various forms of prompt augmentation.[11] Prompting-based personalization techniques, particularly Retrieval-Augmented Generation (RAG), are rapidly evolving to integrate diverse user data into LLMs, offering a flexible and computationally lighter alternative to full model fine-tuning for

dynamic personalization. This approach is highly suitable for real-time applications and adapting to dynamic user preferences, as it bypasses the need for extensive model updates.[38] The advancements in RAG indicate a recognition that continually updating model parameters for every piece of new user information is inefficient. Instead, leveraging external, retrievable memory allows LLMs to access user-specific context without altering their core parametric representations, offering a balance between personalization depth and computational cost.

- **Profile-Augmented Prompting:** These methods explicitly use summarized user preferences and profiles, presented in natural language, to augment the LLM's input at the token level.[11] This can be achieved using non-tuned summarizers, where a frozen LLM directly summarizes user profiles (e.g., Cue-CoT, PAG, ONCE), or tuned summarizers, which are trained to adapt to user preferences and style (e.g., Matryoshka, RewriterSlRl).[11]
- **Retrieval-Augmented Prompting (RAG):** This approach excels at extracting the most relevant records from user data to enhance PLLMs, often by employing an additional external memory. This involves personalized memory construction, which designs mechanisms for retaining and updating memory for efficient retrieval (e.g., non-parametric memory like MemPrompt, TeachMe; or parametric memory like LD-Agent, MemoRAG). It also includes personalized memory retrieval techniques (e.g., LaMP, PEARL, ROPG, HYDRA) that select not only relevant but also representative personalized data.[11] RAG has emerged as a scalable and practical alternative for continual learning, allowing LLMs to adapt to new knowledge by retrieving relevant external information at inference time, rather than modifying the LLM itself.[7] Improvements in encoder models, particularly those leveraging LLM backbones, enhance RAG systems by generating high-quality embeddings that better capture semantic relationships, thereby improving retrieval quality for LLM generation.[7]
- **Soft-Fused Prompting:** This technique differs from profile-augmented prompting by compressing personalized data into soft embeddings, which are generated by a user feature encoder. These soft embeddings can then be integrated into the LLM's input via an input prefix (e.g., UEM, PERSOMA, REGEN, PeaPOD), cross-attention mechanisms (e.g., User-LLM, RECAP), or by directly adjusting the output logits (e.g., GSMN).[11]

This trend towards prompting-based methods points towards hybrid LLM architectures where a base model (parametric memory) is augmented by dynamic, external contextual memory (managed by RAG or similar systems). The primary challenge shifts from training vast models to efficiently managing and retrieving relevant personalized information. This also necessitates robust representation

learning for the external memory (e.g., high-quality embeddings for retrieval [7]) and intelligent memory operations (indexing, retrieval, compression [34]) to ensure effective personalization without noise or excessive overhead.

4.2.2 Model-Level Personalization (Adaptation-based)

This level focuses on designing frameworks to efficiently fine-tune or adapt LLM parameters for personalization, frequently utilizing Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA.[11] The emergence of collaborative PEFT frameworks like Per-Pcs and advanced alignment techniques like P-RLHF signifies a crucial shift towards scalable, privacy-preserving, and fine-grained personalization, directly addressing the practical limitations of one-off fine-tuning. Traditional "one PEFT per user" approaches are often costly and limit communal benefits.[9]

- **One PEFT All Users:** This method involves training on data from all users using a single, shared PEFT module. This can manifest as a single PEFT module (e.g., PLoRA, LM-P, UserIdentifier, Review-LLM) that injects personalized information via user embeddings or IDs, or as a Mixture of Experts (MoE) approach that maintains a set of parallel, independent LoRA weights and employs soft routing to aggregate meta-LoRA weights for more adaptive results (e.g., RecLoRA, iLoRA).[11]
- **One PEFT Per User:** This approach equips each individual user with a specific PEFT module, a strategy that helps preserve data privacy. While some implementations involve no collaboration or coordination between adapters during the learning process for each user (e.g., UserAdapter, PocketLLM, OPPU), others incorporate collaborative efforts to address computational and storage intensity and facilitate knowledge sharing. Examples include PER-PCS, which allows sharing of PEFT parameters, and federated learning frameworks (e.g., Wagner et al., FDLoRA), or models like HYDRA that use a base model with distinct heads for each user.[11]
- **Personalized Pieces (Per-Pcs):** This novel framework enables users to safely share and assemble personalized PEFT efficiently through collaborative efforts.[9] Per-Pcs operates by selecting sharers, breaking their PEFT into smaller "pieces," and training specific "gates" for each piece. These pieces are then added to a shared pool, from which target users can select and assemble personalized PEFT using their own historical data.[9] This modular approach preserves privacy by sharing components rather than entire models or raw data, enables fine-grained user modeling, and significantly reduces storage and computation demands.[9]

Experimental results demonstrate that Per-Pcs outperforms non-personalized and PEFT retrieval baselines, offering performance comparable to other methods with significantly lower resource use across various tasks.[9]

These innovations are critical for scaling personalization to millions of users while effectively managing computational resources and privacy concerns. The modularity of Per-Pcs promotes safe sharing and wider accessibility, potentially fostering community-driven personalization. This also highlights the growing importance of ethical AI design, where privacy preservation and efficient resource use are integrated into the core architectural decisions for personalized LLMs.

4.2.3 Objective-Level Personalization (Alignment-based)

This level focuses on refining LLM behavior to align with individual users' unique preferences, extending beyond generic preferences.[11]

- **Personalized Alignment Data Construction:** High-quality data construction is paramount for effective alignment. This often involves self-generated data derived from interactions with the LLM (e.g., PLUM simulating dynamic interactions, Lee et al. using system messages as meta-instructions).[11] Specialized datasets, such as the PRISM Alignment Dataset and PersonalLLM, are developed to assess the comprehension of personalized preferences.[11]
- **Personalized Alignment Optimization:** This is frequently modeled as a multi-objective reinforcement learning (MORL) problem, where personalized preference is treated as a user-specific combination of multiple preference dimensions.[11] Approaches include using personalized reward models (e.g., MORLHF, MODPO) to guide policy LLMs during training, or ad-hoc combinations of multiple trained policy LLMs during the decoding phase (e.g., Personalized Soups, Reward Soups, MOD).[11] Personalized-RLHF (P-RLHF) is an efficient framework that utilizes a lightweight user model to capture individual user preferences and jointly learns the user model and the personalized LLM from human feedback, without requiring separate reward models for each preference dimension.[14] P-RLHF's ability to handle implicit user preferences from feedback data moves towards more intuitive and less explicit personalization.

4.3 Representation Editing for Fine-Grained Personalization

Representation editing has emerged as a significant technique for model alignment, involving the direct manipulation of a model's latent representations to improve its performance and align it with desired attributes.[40] While this technique has been recognized and applied in visual generation models, its exploration for personalizing LLMs has been relatively limited until recently.[40]

A notable contribution in this area is the work "Personalize Your LLM: Fake it then Align it" (NAACL Findings 2025). This research proposes a scalable and efficient personalization approach that leverages self-generated personal preference data and representation editing to enable quick and cost-effective personalization.[40] The method specifically focuses on identifying embedding spaces that capture personalized versus non-personalized preferences and then performs personalization by directly editing these representations.[40] This approach represents an emerging frontier for LLM personalization, offering a precise and efficient method to align models with user preferences by directly manipulating latent representations, moving beyond traditional fine-tuning. Unlike broad fine-tuning, representation editing allows for a more granular and potentially more efficient way to steer LLM behavior towards specific user preferences (e.g., style, tone, factual alignment) without retraining the entire model or large portions of it. This implies a deeper understanding of the LLM's internal "thought process" (its latent space) and the ability to surgically alter it. This could lead to more interpretable personalized LLMs, as researchers can directly observe how changes in the embedding space correlate with changes in personalized output. It also offers greater control over personalization, potentially mitigating issues like bias or unintended side effects that can arise from less targeted fine-tuning, which is particularly valuable for sensitive applications where precise control over LLM behavior is paramount.

5. Practical Applications and Case Studies of Personalized LLMs

5.1 Conversational AI and Intelligent Agents

LLMs are fundamentally transforming the landscape of conversational AI and intelligent agents, enabling seamless and contextually relevant dialogues across diverse topics.[12] Personalized LLM agents are demonstrating the capacity to adapt to individual user needs, fostering emotional bonds and encouraging sustained interactions.[12] LLM-powered chatbots are evolving from basic conversational tools into sophisticated assistants capable of understanding nuanced context and delivering highly personalized recommendations. These advanced chatbots streamline customer support by accurately interpreting complex customer queries, minimizing the need for human intervention, and providing faster responses. They also proactively engage with users, for instance, by addressing abandoned carts in e-commerce scenarios.[43]

In the healthcare sector, generative LLMs are being deployed for critical applications such as real-time, no-code COVID-19 severity prediction through conversational, question-answering interactions. These models have demonstrated strong performance even in low-data settings, showcasing their adaptability to dynamic clinical environments.[44] The integration of representation learning and personalization in conversational AI is moving beyond basic chatbot functionalities to create emotionally resonant and domain-specific intelligent agents that significantly enhance user engagement and provide critical real-time support. The shift from generic responses to personalized ones, enabled by improved representation of user preferences and context, transforms the user experience from mere utility to a more engaging and even empathetic interaction. In healthcare, this means a transition from static diagnostic tools to dynamic, conversational AI that can adapt to individual patient data, providing timely and tailored risk assessments. This represents a significant leap from purely functional AI to emotionally and contextually intelligent AI. Enhanced personalization also fosters greater user trust and adoption. However, it also raises ethical considerations around data privacy (especially sensitive health data), the potential for over-reliance, and the imperative for robust mechanisms to ensure accuracy and prevent harmful outputs, particularly in critical applications like healthcare. The ability to handle "low-data settings" [44] is a key representation learning improvement for real-world deployment.

5.2 Revolutionizing Recommendation Systems

LLMs are introducing a new paradigm for recommender systems, enhancing personalization, semantic alignment, and interpretability without requiring extensive

task-specific supervision.[45] Their inherent capabilities enable zero- and few-shot reasoning, allowing these systems to operate effectively even in challenging cold-start and long-tail scenarios where historical data is scarce.[45] LLMs are not merely auxiliary components in recommendation systems; they are foundational enablers for constructing more adaptive, semantically rich, and user-centric systems.[45]

The Memory-Assisted Personalized LLM (MAP) framework exemplifies these advancements. MAP leverages user interactions to construct detailed history profiles, capturing individual preferences such as item ratings. When a recommendation is requested, MAP's retrieval module selectively extracts the most relevant data points from the user profile based on the current query. This is achieved by computing a similarity score between the item to be predicted and the historical items stored in the user's memory. For instance, for items with genre lists, similarity can be calculated by comparing genre intersections. A more advanced technique involves using pre-trained language models like BERT for text feature extraction, embedding text descriptions of both historical and predicted items, and computing cosine similarity between their feature vectors. This allows for semantic comparison, identifying similarities even if different terms are used in descriptions.[38] Once the most relevant items are identified and ranked based on similarity, this processed memory is integrated into the LLM as part of the prompt, enabling the LLM to generate more personalized and contextually accurate recommendations by focusing on the user's most relevant historical preferences, while also reducing computational costs by limiting the input to only pertinent data.[38] Experimental results indicate that MAP consistently outperforms regular LLM-based recommenders that directly integrate user history through prompt design, with its advantage becoming more pronounced as user history grows, making it highly suitable for addressing successive personalized user requests.[38]

Another innovative approach is the Rec4Agentverse, which proposes a new recommendation paradigm built on an LLM-based agent platform. This framework offers personalized suggestions from specialized "Item Agents" to users via an "Agent Recommender." These Item Agents can gather user data through direct interactions or by accessing relevant records, and they can collaborate with other agents to gain broader insights into user preferences, leading to more flexible and tailored recommendations.[46] Personalized LLMs are fundamentally reshaping recommendation systems by enabling deeper semantic understanding of user preferences and items, facilitating effective cold-start and long-tail recommendations, and integrating dynamic memory for continuous adaptation. This advancement enables hyper-personalized shopping experiences [43], potentially increasing conversions and customer retention. However, it also introduces challenges related to fairness and bias

(LLMs may amplify biases from training data), privacy (users may inadvertently disclose private information, especially in multi-agent collaboration), and harmfulness (agents generating harmful responses or transactions).[46] Addressing these requires careful design of prompt injection to reduce bias, user control over data access, and robust admission mechanisms for agents.

5.3 Domain-Specific Implementations (E-commerce, Education)

The successful deployment of personalized LLMs across diverse domains like e-commerce and education demonstrates their generalizability and the critical role of adaptive representation learning in tailoring AI solutions to specific industry needs. The ability of LLMs to excel in these domains is a testament to their underlying representation learning capabilities, which allow them to capture intricate patterns from vast general datasets and then adapt effectively to domain-specific nuances through fine-tuning or prompting. This adaptation creates significant value.

- **E-commerce:** LLMs are significantly enhancing product search optimization within e-commerce platforms. They achieve this by grasping the semantic meanings in natural language queries, utilizing synonyms, spell corrections, and relaxation rules.[43] This leads to superior handling of ambiguous or long-tail searches (e.g., "running shoes good for knees") and improved interpretation of typos, ensuring customers are connected with relevant products even when their queries are imprecise.[43] Beyond search, LLMs are also instrumental in generating high-quality product descriptions and SEO-optimized content, which substantially increases online visibility and attracts more qualified traffic to e-commerce platforms.[43] This translates into increased conversions through "hyper-personalized shopping experiences".[43]
- **Education:** In the educational sector, LLMs hold immense potential to enhance pedagogical efficacy and are increasingly perceived as valuable tools. They serve as effective teaching assistants, foster creativity among learners, democratize access to educational technology, and adapt to specific educational contexts.[47] Personalized LLMs can provide tailored learning plans and support, directly addressing individual student needs, knowledge gaps, or motivational challenges.[48] This includes capabilities such as adaptive content delivery, real-time feedback, and the development of intelligent tutoring systems that cater to diverse learning styles and paces.[48] The "no-code" aspect mentioned for healthcare applications [44] also points to the increasing ease of deployment,

making these powerful tools accessible to non-technical domain experts in education.

6. Key Challenges and Future Research Directions

6.1 Technical Hurdles: Efficiency, Data Sparsity, Overfitting, Complex Data Integration

Despite the rapid advancements, the personalization of LLMs faces several persistent technical hurdles. The primary technical challenges revolve around balancing computational efficiency with the depth and dynamism of user representation, particularly in managing memory and integrating diverse, sparse, and evolving user data across different deployment environments. This presents a paradox: true personalization requires rich, dynamic, and continuously updated user representations, but achieving this is computationally intensive and prone to issues with sparse, real-world user data. The challenge is not just *how* to personalize, but *how to personalize efficiently, robustly, and continuously* across heterogeneous data types and deployment scenarios.

- **Efficiency:** LLMs demand substantial computational resources for both training and inference, resulting in high costs and latency. This makes them less practical for real-time applications or environments with limited resources, such as edge devices.[1]
- **Data Sparsity & Overfitting:** Fine-tuning methods are susceptible to overfitting, especially when working with limited or noisy user data, which can compromise generalization capabilities. A significant challenge lies in constructing high-quality, user-specific preference datasets due to inherent data sparsity.[11]
- **Complex Data Integration:** Efficiently representing and integrating diverse user data, including profiles, historical dialogues, content, and interactions, remains a complex task.[11] Moreover, effectively leveraging complex, multi-source user information, such as user relationships in graph-like structures, to fine-tune LLM parameters is still difficult. Most current methods primarily focus on text data, leaving personalized foundation models for multimodal data largely underexplored.[11]

- **Memory Management:** Retaining excessive information can lead to noisy memory representations, inefficient retrieval processes, and an increased propensity for LLM hallucinations. Conversely, storing too little information can compromise performance. The issue of forgetting information is exacerbated as context windows continue to expand.[2] Preventing catastrophic forgetting while ensuring the efficient updating of both long-term and short-term memory is a crucial ongoing challenge.[7]
- **Edge Computing:** Efficiently updating models on resource-constrained devices, such as smartphones, and ensuring seamless synchronization between cloud and edge devices in real-world deployments, presents a significant technical barrier.[11]

Overcoming these challenges necessitates a convergence of research in various areas, including representation learning (e.g., efficient embeddings for sparse data), memory management (e.g., selective forgetting, intelligent indexing), model compression (e.g., quantization for edge devices), and data curation (e.g., synthetic data generation for alignment). This indicates that future breakthroughs will likely stem from interdisciplinary approaches that integrate solutions from these seemingly disparate fields.

6.2 Ethical Considerations: Privacy, Bias, and Trustworthiness

Ethical considerations, particularly privacy, bias, and trustworthiness, are not secondary concerns but fundamental design constraints for personalized LLMs, requiring the proactive integration of responsible AI principles into representation learning and personalization techniques. These issues are not just post-deployment problems; they impact the core design of how representations are learned and how personalization is achieved.

- **Privacy:** The sharing of personal data raises significant concerns regarding its storage, usage, and protection.[11] Users may inadvertently disclose private information, especially during collaborative interactions with agents, where sensitive data might circulate among multiple entities.[46] Ensuring users maintain control over their privacy and implementing data minimization principles are critical safeguards.[46] This drives the development of techniques like Per-Pcs [9] that share "pieces" instead of raw data, or focus on on-device processing.[46]
- **Bias:** LLMs have the potential to learn, perpetuate, and amplify harmful social biases present in their training data, which can lead to unfair or discriminatory

outcomes.[12] It is imperative to acknowledge and actively control potential unfairness and bias in recommended agents and the information they provide, possibly through carefully designed prompt injection or rigorous output checks.[46] Bias mitigation requires careful data curation [1] and representation editing [40] to align models with desired values.

- **Trustworthiness/Hallucinations:** LLMs can produce content that is factually incorrect or contextually inappropriate, a phenomenon known as "hallucination," often due to limitations in training data or algorithmic flaws.[12] The "black box" nature of many LLMs also raises concerns regarding their transparency and interpretability.[12] Additionally, there is a risk of harmful textual responses or manipulated actions being executed by LLM-based agents.[46] Trustworthiness necessitates explainable models [6] and robust evaluation.

The pervasive nature of personalized LLMs means these ethical challenges have significant societal implications, affecting inclusivity, fairness, and public trust. This will likely lead to increased regulatory scrutiny and the development of industry standards for responsible AI, influencing research funding and priorities towards ethical considerations in representation learning and personalization. The "black box" nature [12] also means that interpretability of representations becomes a key research area.

6.3 The Pursuit of Lifelong Learning and Dynamic Adaptation

The ultimate goal for personalized LLMs is to achieve true lifelong learning and dynamic adaptation. This entails models continuously evolving their representations and personalization strategies in real-time based on ongoing user interactions, without forgetting previously acquired knowledge. The ability to acquire and integrate new knowledge over time while preserving past information is crucial for LLMs in real-world applications.[7] This involves a combination of techniques, including continual fine-tuning, model editing, and Retrieval-Augmented Generation (RAG).[7]

Ensuring adaptivity to accommodate users' diverse and evolving needs and behaviors, while preventing catastrophic forgetting, are central challenges for managing long-term user memories.[2] Current LLMs, even personalized ones, are often static snapshots of knowledge. True lifelong learning implies a dynamic system where representations are not fixed but constantly refined based on new interactions and evolving user preferences. This moves beyond discrete fine-tuning cycles to a

continuous learning process, which is essential for personal assistants or domain-specific agents that need to maintain a "running memory" of user interactions.[2] This requires sophisticated memory management and adaptive learning algorithms that can integrate new information without compromising existing knowledge. Achieving lifelong learning would enable a more symbiotic relationship between humans and AI, where the AI truly "understands" and adapts to the individual over extended periods. This is a critical step towards highly autonomous and intelligent agents that can learn from experience, self-improve [36], and provide truly personalized, evolving support across complex, multi-session interactions. This will necessitate advancements in areas like meta-learning, transfer learning, and robust memory architectures.

7. Conclusion

The rapid evolution of Large Language Models has established them as a transformative force in artificial intelligence, with representation learning and personalization emerging as intertwined and critical areas for their continued advancement. The exponential growth in research, coupled with a notable shift in leadership from industry to academia, underscores the field's dynamism and increasing accessibility. Significant progress has been made in enhancing LLM efficiency through the development of smaller models and techniques like PEFT and quantization, addressing the substantial computational overhead of larger models. Concurrently, data-centric approaches are refining representation quality by prioritizing curated, high-quality datasets and leveraging LLMs themselves for synthetic data generation.

A deeper understanding and operationalization of memory mechanisms, distinguishing between parametric and contextual forms and defining fundamental operations, are proving essential for building robust and dynamically adaptive personalized LLMs. The expansion into multimodal representation learning further promises more comprehensive, human-like understanding, profoundly impacting how personalization can be achieved across diverse interaction modalities.

Personalization itself has matured into a sophisticated domain, with a comprehensive taxonomy of techniques ranging from flexible input-level prompting (e.g., advanced RAG) to efficient model-level adaptation (e.g., collaborative PEFT frameworks like Per-Pcs) and objective-level alignment (e.g., P-RLHF and representation editing).

These innovations are paving the way for scalable, privacy-preserving, and fine-grained personalization, moving beyond the limitations of traditional fine-tuning.

The practical impact of personalized LLMs is already evident across various sectors. In conversational AI, they are enabling emotionally resonant and domain-specific intelligent agents that enhance user engagement and provide critical real-time support, as seen in healthcare applications. In recommendation systems, personalized LLMs are revolutionizing the field by enabling deeper semantic understanding, facilitating effective cold-start and long-tail recommendations, and integrating dynamic memory for continuous adaptation. Their successful deployment in e-commerce and education further demonstrates their generalizability and the crucial role of adaptive representation learning in tailoring AI solutions to specific industry needs.

However, the path forward is not without significant challenges. Technical hurdles persist in balancing computational efficiency with the depth and dynamism of user representation, particularly in managing memory and integrating diverse, sparse, and evolving user data across different deployment environments. Equally critical are the ethical considerations surrounding privacy, bias, and trustworthiness. These are not merely secondary concerns but fundamental design constraints that necessitate the proactive integration of responsible AI principles into every stage of representation learning and personalization.

The future of LLMs lies in achieving true lifelong learning and dynamic adaptation, where models continuously evolve their representations and personalization strategies in real-time based on ongoing user interactions, without forgetting past knowledge. This pursuit promises a more symbiotic relationship between humans and AI, leading to highly autonomous and intelligent agents that can learn from experience, self-improve, and provide truly personalized, evolving support across complex, multi-session interactions. Continued research in these areas will ensure that LLMs are not just powerful tools, but also deeply personal, adaptive, and responsibly integrated into the fabric of human lives.

References

- [15] https://nlplab-skku.github.io/Publications/UpcomingConferences/?ref=thefragiles

ea.com
- [16] https://www.datacamp.com/blog/top-ai-conferences
- [17] https://www.google.com/search?q=top+journals+natural+language+processing
- [18] https://www.google.com/search?q=top+journals+machine+learning
- [1] https://arxiv.org/pdf/2409.06857
- [5] https://arxiv.org/html/2504.08619v1
- [8] https://arxiv.org/abs/2411.00027
- [9] https://aclanthology.org/2024.emnlp-main.371/
- [37] https://github.com/pliang279/awesome-multimodal-ml
- [35] https://arxiv.org/html/2505.00675v2
- [10] https://www.researchgate.net/publication/385510317_Personalization_of_Large_Language_Models_A_Survey
- [11] https://arxiv.org/abs/2502.11528
- [33] https://openreview.net/forum?id=TJENqwWsq6
- [34] https://arxiv.org/html/2505.00675v1
- [3] https://arxiv.org/pdf/2504.08619
- [5] https://arxiv.org/html/2504.08619v1
- [7] https://icml.cc/virtual/2025/poster/45585
- [34] https://arxiv.org/html/2505.00675v1
- [32] https://github.com/HuangOwen/Awesome-LLM-Compression
- [2]

https://arxiv.org/html/2507.03042v1
- [36] https://github.com/Elvin-Yiming-Du/Survey_Memory_in_AI
- [45] https://arxiv.org/html/2507.21117v1
- [1] https://arxiv.org/pdf/2409.06857
- [6] https://www.cis.upenn.edu/~ccb/research-statement
- [23] https://www.oii.ox.ac.uk/news-events/large-language-models-llms-getting-personal/
- [45] https://arxiv.org/html/2507.21117v1
- [4] https://arxiv.org/html/2501.10326v2
- [44] https://ai.jmir.org/2025/1/e67363
- [43] https://www.netguru.com/blog/llm-use-cases-in-e-commerce
- [39] https://www.themoonlight.io/en/review/on-the-way-to-llm-personalization-learning-to-remember-user-conversations
- [40] https://aclanthology.org/2025.findings-naacl.407.pdf
- [44] https://ai.jmir.org/2025/1/e67363
- [12] https://arxiv.org/html/2402.15265v1
- [13] https://arxiv.org/abs/2412.11736
- [30] https://github.com/TamSiuhin/Per-Pcs
- [31] https://arxiv.org/abs/2406.10471
- [42] https://aclanthology.org/2025.findings-naacl.407/
- [41] https://arxiv.org/pdf/2503.01048

- [46] https://cacm.acm.org/research/envisioning-recommendations-on-an-llm-based-agent-platform/
- [38] https://www.arxiv.org/pdf/2505.03824
- [44] https://ai.jmir.org/2025/1/e67363
- [49] https://addepto.com/blog/llm-use-cases-for-business/
- [47] https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1392091/full
- [48] https://www.researchgate.net/publication/391151729_LearnMate_Enhancing_Online_Education_with_LLM-Powered_Personalized_Learning_Plans_and_Support
- [14] https://arxiv.org/pdf/2402.05133
- [29] https://machinelearning.apple.com/research/on-the-way
- [50] https://hugolab.wustl.edu/people/zhehao-zhang/
- [51] https://saltlab.stanford.edu/authors/zhehao-zhang/
- [27] https://research.adobe.com/person/ryan-rossi/
- [28] http://ryanrossi.com/
- [19] https://cs.stanford.edu/~diyiy/group.html
- [20] https://cs.stanford.edu/~diyiy/
- [24] https://groups.cs.umass.edu/zamani/research/
- [25] https://groups.cs.umass.edu/zamani/
- [52] https://harpercancer.nd.edu/people/meng-jiang/
- [53] https://engineering.nd.edu/faculty/meng-jiang/

- [21] https://pages.cs.wisc.edu/~fredsala/
- [22] https://snorkel.ai/author/fred-sala/
- [54] https://www.kw.lmu.de/dch/en/institute/team/contact-page/yiming-du-a3ed9166.html
- [26] https://dblp.org/pid/230/6249
- [9] https://aclanthology.org/2024.emnlp-main.371/
- [11] https://arxiv.org/abs/2502.11528
- [3] https://arxiv.org/pdf/2504.08619
- [23] https://www.oii.ox.ac.uk/news-events/large-language-models-llms-getting-personal/
- [38] https://www.arxiv.org/pdf/2505.03824
- [51] https://saltlab.stanford.edu/authors/zhehao-zhang/
- [27] https://research.adobe.com/person/ryan-rossi/
- [24] https://groups.cs.umass.edu/zamani/research/
- [21] https://pages.cs.wisc.edu/~fredsala/
- [26] https://dblp.org/pid/230/6249
- [29] https://machinelearning.apple.com/research/on-the-way
- [23] https://www.oii.ox.ac.uk/news-events/large-language-models-llms-getting-personal/

Works cited

1. arXiv:2409.06857v5 [cs.CL] 15 Apr 2025, accessed July 30, 2025, https://arxiv.org/pdf/2409.06857

2. Dynamic LSTM-based Memory Encoder For Long-term LLM Interactions - arXiv, accessed July 30, 2025, https://arxiv.org/html/2507.03042v1
3. Analyzing 16193 LLM Papers for Fun and Profits - arXiv, accessed July 30, 2025, https://arxiv.org/pdf/2504.08619
4. Large language models for automated scholarly paper review: A survey - arXiv, accessed July 30, 2025, https://arxiv.org/html/2501.10326v2
5. arxiv.org, accessed July 30, 2025, https://arxiv.org/html/2504.08619v1
6. Research Statement - Chris Callison-Burch - CIS UPenn - University of Pennsylvania, accessed July 30, 2025, https://www.cis.upenn.edu/~ccb/research-statement
7. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models - ICML 2025, accessed July 30, 2025, https://icml.cc/virtual/2025/poster/45585
8. [2411.00027] Personalization of Large Language Models: A Survey - arXiv, accessed July 30, 2025, https://arxiv.org/abs/2411.00027
9. Personalized Pieces: Efficient Personalized Large Language ..., accessed July 30, 2025, https://aclanthology.org/2024.emnlp-main.371/
10. (PDF) Personalization of Large Language Models: A Survey, accessed July 30, 2025, https://www.researchgate.net/publication/385510317_Personalization_of_Large_Language_Models_A_Survey
11. A Survey of Personalized Large Language Models: Progress and ..., accessed July 30, 2025, https://arxiv.org/abs/2502.11528
12. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models - arXiv, accessed July 30, 2025, https://arxiv.org/html/2402.15265v1
13. [2412.11736] Personalized LLM for Generating Customized Responses to the Same Query from Different Users - arXiv, accessed July 30, 2025, https://arxiv.org/abs/2412.11736
14. personalized language modeling from personalized human feedback - arXiv, accessed July 30, 2025, https://arxiv.org/pdf/2402.05133
15. Upcoming Conferences - NLPLAB, accessed July 30, 2025, https://nlplab-skku.github.io/Publications/UpcomingConferences/?ref=thefragileseea.com
16. Top 10 AI Conferences for 2025 | DataCamp, accessed July 30, 2025, https://www.datacamp.com/blog/top-ai-conferences
17. www.google.com, accessed July 30, 2025, https://www.google.com/search?q=top+journals+natural+language+processing
18. www.google.com, accessed July 30, 2025, https://www.google.com/search?q=top+journals+machine+learning
19. Diyi Yang - Stanford Computer Science, accessed July 30, 2025, https://cs.stanford.edu/~diyiy/group.html
20. Diyi Yang - CS Stanford, accessed July 30, 2025, https://cs.stanford.edu/~diyiy/
21. Frederic Sala, University of Wisconsin-Madison - cs.wisc.edu, accessed July 30, 2025, https://pages.cs.wisc.edu/~fredsala/

22. Fred Sala | Snorkel AI, accessed July 30, 2025, https://snorkel.ai/author/fred-sala/
23. Large Language Models (LLMs): Getting personal - OII, accessed July 30, 2025, https://www.oii.ox.ac.uk/news-events/large-language-models-llms-getting-personal/
24. Research – Hamed Zamani - UMass Amherst, accessed July 30, 2025, https://groups.cs.umass.edu/zamani/research/
25. Hamed Zamani – Associate Professor at the Manning College of Information and Computer Sciences - UMass Amherst, accessed July 30, 2025, https://groups.cs.umass.edu/zamani/
26. Yiming Du - dblp, accessed July 30, 2025, https://dblp.org/pid/230/6249
27. Adobe Research » Ryan A. Rossi, accessed July 30, 2025, https://research.adobe.com/person/ryan-rossi/
28. Ryan A. Rossi - Ph.D., accessed July 30, 2025, http://ryanrossi.com/
29. On the Way to LLM Personalization: Learning to Remember User ..., accessed July 30, 2025, https://machinelearning.apple.com/research/on-the-way
30. TamSiuhin/Per-Pcs: Official Implementation of "Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts" at EMNLP 2024 Main Conference - GitHub, accessed July 30, 2025, https://github.com/TamSiuhin/Per-Pcs
31. [2406.10471] Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts - arXiv, accessed July 30, 2025, https://arxiv.org/abs/2406.10471
32. HuangOwen/Awesome-LLM-Compression - GitHub, accessed July 30, 2025, https://github.com/HuangOwen/Awesome-LLM-Compression
33. Rethinking Memory in AI: Taxonomy, Operations, Topics, and Future ..., accessed July 30, 2025, https://openreview.net/forum?id=TJENqwWsq6
34. Rethinking Memory in AI: Taxonomy, Operations, Topics, and Future Directions - arXiv, accessed July 30, 2025, https://arxiv.org/html/2505.00675v1
35. Rethinking Memory in AI: Taxonomy, Operations, Topics, and Future Directions - arXiv, accessed July 30, 2025, https://arxiv.org/html/2505.00675v2
36. Elvin-Yiming-Du/Survey_Memory_in_AI: This repository introduce a comprehensive paper list, datasets, methods and tools for memory research. - GitHub, accessed July 30, 2025, https://github.com/Elvin-Yiming-Du/Survey_Memory_in_AI
37. pliang279/awesome-multimodal-ml: Reading list for research topics in multimodal machine learning - GitHub, accessed July 30, 2025, https://github.com/pliang279/awesome-multimodal-ml
38. Memory Assisted LLM for Personalized Recommendation ... - arXiv, accessed July 30, 2025, https://www.arxiv.org/pdf/2505.03824
39. [Literature Review] On the Way to LLM Personalization: Learning to Remember User Conversations - Moonlight, accessed July 30, 2025, https://www.themoonlight.io/en/review/on-the-way-to-llm-personalization-learning-to-remember-user-conversations
40. Personalize Your LLM: Fake it then Align it - ACL Anthology, accessed July 30, 2025, https://aclanthology.org/2025.findings-naacl.407.pdf

41. arXiv:2503.01048v3 [cs.LG] 5 Mar 2025, accessed July 30, 2025, https://arxiv.org/pdf/2503.01048
42. Personalize Your LLM: Fake it then Align it - ACL Anthology, accessed July 30, 2025, https://aclanthology.org/2025.findings-naacl.407/
43. 17 Proven LLM Use Cases in E-commerce That Boost Sales in 2025, accessed July 30, 2025, https://www.netguru.com/blog/llm-use-cases-in-e-commerce
44. Generative Large Language Model—Powered ... - JMIR AI, accessed July 30, 2025, https://ai.jmir.org/2025/1/e67363
45. A Comprehensive Review on Harnessing Large Language Models to Overcome Recommender System Challenges - arXiv, accessed July 30, 2025, https://arxiv.org/html/2507.21117v1
46. Envisioning Recommendations on an LLM-Based Agent Platform, accessed July 30, 2025, https://cacm.acm.org/research/envisioning-recommendations-on-an-llm-based-agent-platform/
47. The impact of large language models on higher education: exploring the connection between AI and Education 4.0 - Frontiers, accessed July 30, 2025, https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1392091/full
48. LearnMate: Enhancing Online Education with LLM-Powered Personalized Learning Plans and Support | Request PDF - ResearchGate, accessed July 30, 2025, https://www.researchgate.net/publication/391151729_LearnMate_Enhancing_Online_Education_with_LLM-Powered_Personalized_Learning_Plans_and_Support
49. 15 LLM Use Cases in 2025: Integrate LLM Models to Your Business - Addepto, accessed July 30, 2025, https://addepto.com/blog/llm-use-cases-for-business/
50. Zhehao Zhang | Computational Radiotherapy Lab | Washington University in St. Louis, accessed July 30, 2025, https://hugolab.wustl.edu/people/zhehao-zhang/
51. Zhehao Zhang | Stanford SALT Lab, accessed July 30, 2025, https://saltlab.stanford.edu/authors/zhehao-zhang/
52. Meng - Jiang // People // Mike and Josie Harper Cancer Research Institute // University of Notre Dame, accessed July 30, 2025, https://harpercancer.nd.edu/people/meng-jiang/
53. Meng Jiang - College of Engineering, accessed July 30, 2025, https://engineering.nd.edu/faculty/meng-jiang/
54. Contact page - Digital Cultural Heritage Studies - LMU München, accessed July 30, 2025, https://www.kw.lmu.de/dch/en/institute/team/contact-page/yiming-du-a3ed9166.html