Technical Review: LoRA for Smartphone

Introduction

This document provides a technical review of research papers focusing on the application of Low-Rank Adaptation (LoRA) techniques to smartphone platforms. LoRA's ability to efficiently fine-tune large language models (LLMs) makes it a promising area for on-device AI, addressing limitations of computational power and memory in smartphones.

Search Results

Paper 1: K-LoRA: Unlocking Training-Free Fusion of Any Subject and Style LoRAs

Authors: Ziheng Ouyang, Zhen Li, Qibin Hou

Published: 2025-02-25

Summary: Recent studies have explored combining different LoRAs to jointly generate learned style and content. However, existing methods either fail to effectively preserve both the original subject and style simultaneously or require additional training. In this paper, we argue that the intrinsic properties of LoRA can effectively guide diffusion models in merging learned subject and style. Building on this insight, we propose K-LoRA, a simple yet effective training-free LoRA fusion approach. In each attention layer, K-LoRA compares the Top-K elements in each LoRA to be fused, determining which LoRA to select for optimal fusion. This selection mechanism ensures that the most representative features of both subject and style are retained during the fusion process, effectively balancing their contributions. Experimental results demonstrate that the proposed method effectively integrates the subject and style information learned by the original LoRAs, outperforming state-of-the-art training-based approaches in both qualitative and quantitative results.

arXiv ID: 2502.18461v1

URL: K-LoRA: Unlocking Training-Free Fusion of Any Subject and Style LoRAs

---

Paper 2: DRAMA: Diverse Augmentation from Large Language Models to Smaller Dense Retrievers

Authors: Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, Xilun Chen

Published: 2025-02-25

Summary: Large language models (LLMs) have demonstrated strong effectiveness and robustness while fine-tuned as dense retrievers. However, their large parameter size brings significant inference time computational challenges, including high encoding costs for large-scale corpora and increased query latency, limiting their practical deployment. While smaller retrievers offer better efficiency, they often fail to generalize effectively with limited supervised fine-tuning data. In this work, we introduce DRAMA, a training framework that leverages LLMs to train smaller generalizable dense retrievers. In particular, we adopt pruned LLMs as the backbone and train on diverse LLM-augmented data in a single-stage contrastive learning setup. Experiments show that DRAMA offers better multilingual and long-context capabilities than traditional encoder-based retrievers, and achieves strong performance across multiple tasks and languages. These highlight the potential of connecting the training of smaller retrievers with the growing advancements in LLMs, bridging the gap between efficiency and generalization.

arXiv ID: 2502.18460v1

URL: DRAMA: Diverse Augmentation from Large Language Models to Smaller Dense Retrievers

---

Paper 3: LLM-Based Design Pattern Detection

Authors: Christian Schindler, Andreas Rausch

Published: 2025-02-25

Summary: Detecting design pattern instances in unfamiliar codebases remains a challenging yet essential task for improving software quality and maintainability. Traditional static analysis tools often struggle with the complexity, variability, and lack of explicit annotations that characterize real-world pattern implementations. In this paper, we present a novel approach leveraging Large Language Models to automatically identify design pattern instances across diverse codebases. Our method focuses on recognizing the roles classes play within the pattern instances. By providing clearer insights into software structure and intent, this research aims to support developers, improve comprehension, and streamline tasks such as refactoring, maintenance, and adherence to best practices.

arXiv ID: 2502.18458v1

URL: LLM-Based Design Pattern Detection

Paper 4: FRIDA to the Rescue! Analyzing Synthetic Data Effectiveness in Object-Based Common Sense Reasoning for Disaster Response

Authors: Mollie Shichman, Claire Bonial, Austin Blodgett, Taylor Hudson, Francis Ferraro, Rachel Rudinger

Published: 2025-02-25

Summary: Large Language Models (LLMs) have the potential for substantial common sense reasoning. However, these capabilities are often emergent in larger models. This means smaller models that can be run locally are less helpful and capable with respect to certain reasoning tasks. To meet our problem space requirements, we fine-tune smaller LLMs to disaster domains, as these domains involve complex and low-frequency physical common sense knowledge. We introduce a pipeline to create Field Ready Instruction Decoding Agent (FRIDA) models, where domain experts and linguists combine their knowledge to make high-quality seed data that is used to generate synthetic data for fine-tuning. We create a set of 130 seed instructions for synthetic generation, a synthetic dataset of 25000 instructions, and 119 evaluation instructions relating to both general and earthquake-specific object affordances. We fine-tune several LLaMa and Mistral instruction-tuned models and find that FRIDA models outperform their base models at a variety of sizes. We then run an ablation study to understand which kinds of synthetic data most affect performance and find that training physical state and object function common sense knowledge alone improves over FRIDA models trained on all data. We conclude that the FRIDA pipeline is capable of instilling general common sense, but needs to be augmented with information retrieval for specific domain knowledge.

arXiv ID: 2502.18452v1

URL: [FRIDA to the Rescue! Analyzing Synthetic Data Effectiveness in Object-Based Common Sense Reasoning for Disaster Response]

Paper 5: SWE-RL: Advancing LLM Reasoning via Reinforcement Learning on Open Software Evolution

Authors: Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, Sida I. Wang

Published: 2025-02-25

Summary: The recent DeepSeek-R1 release has demonstrated the immense potential of reinforcement learning (RL) in enhancing the general reasoning capabilities of large language models (LLMs). While DeepSeek-R1 and other follow-up work primarily focus on applying RL to competitive coding and math problems, this paper introduces SWE-RL, the first approach to scale RL-based LLM reasoning for real-world software engineering. Leveraging a lightweight rule-based reward (e.g., the similarity score between ground-truth and LLM-generated solutions), SWE-RL enables LLMs to autonomously recover a developer's reasoning processes and solutions by learning from extensive open-source software evolution data -- the record of a software's entire lifecycle, including its code snapshots, code changes, and events such as issues and pull requests. Trained on top of Llama 3, our resulting reasoning model, Llama3-SWE-RL-70B, achieves a 41.0% solve rate on SWE-bench Verified -- a human-verified collection of real-world GitHub issues. To our knowledge, this is the best performance reported for medium-sized (<100B) LLMs to date, even comparable to leading proprietary LLMs like GPT-4o. Surprisingly, despite performing RL solely on software evolution data, Llama3-SWE-RL has even emerged with generalized reasoning skills. For example, it shows improved results on five out-of-domain tasks, namely, function coding, library use, code reasoning, mathematics, and general language understanding, whereas a supervised-finetuning baseline even leads to performance degradation on average. Overall, SWE-RL opens up a new direction to improve the reasoning capabilities of LLMs through reinforcement learning on massive software engineering data.

arXiv ID: 2502.18449v1

URL: [SWE-RL: Advancing LLM Reasoning via Reinforcement Learning on Open Software Evolution]

Paper 6: Disambiguate First Parse Later: Generating Interpretations for Ambiguity Resolution in Semantic Parsing

Authors: Irina Saparina, Mirella Lapata

Published: 2025-02-25

Summary: Handling ambiguity and underspecification is an important challenge in natural language interfaces, particularly for tasks like text-to-SQL semantic parsing. We propose a modular approach that resolves ambiguity using natural language interpretations before mapping these to logical forms (e.g., SQL queries). Although LLMs excel at parsing unambiguous utterances, they show strong biases for ambiguous ones, typically predicting only preferred interpretations. We constructively exploit this bias to generate an initial set of preferred disambiguations and then apply a specialized infilling model to identify and generate missing interpretations. To train the infilling model, we introduce an annotation method that uses SQL execution to validate different meanings. Our approach improves interpretation coverage and generalizes across datasets with different annotation styles, database structures, and ambiguity types.

arXiv ID: 2502.18448v1

URL: Disambiguate First Parse Later: Generating Interpretations for Ambiguity Resolution in Semantic Parsing

---

Paper 7: MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning

Authors: Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman Ozdaglar, Kaiqing Zhang, Joo-Kyung Kim

Published: 2025-02-25

Summary: Leveraging multiple large language models (LLMs) to build collaborative multi-agentic workflows has demonstrated significant potential. However, most previous studies focus on prompting the out-of-the-box LLMs, relying on their innate capability for collaboration, which may not improve LLMs' performance as shown recently. In this paper, we introduce a new post-training paradigm MAPoRL (Multi-Agent Post-co-training for collaborative LLMs with Reinforcement Learning), to explicitly elicit the collaborative behaviors and further unleash the power of multi-agentic LLM frameworks. In MAPoRL, multiple LLMs first generate their own responses independently and engage in a multi-turn discussion to collaboratively improve the final answer. In the end, a MAPoRL verifier evaluates both the answer and the discussion, by assigning a score that verifies the correctness of the answer, while adding incentives to encourage corrective and persuasive discussions. The score serves as the co-training reward, and is then maximized through multi-agent RL. Unlike existing LLM post-training paradigms, MAPoRL advocates the co-training of multiple LLMs together using RL for better generalization. Accompanied by analytical insights, our experiments demonstrate that training individual LLMs alone is insufficient to induce effective collaboration. In contrast, multi-agent co-training can boost the collaboration performance across benchmarks, with generalization to unseen domains.

arXiv ID: 2502.18439v1

URL: MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning

---

Paper 8: Reversal Blessing: Thinking Backward May Outpace Thinking Forward in Multi-choice Questions

Authors: Yizhe Zhang, Richard Bai, Zijin Gu, Ruixiang Zhang, Jiatao Gu, Emmanuel Abbe, Samy Bengio, Navdeep Jaitly

Published: 2025-02-25

Summary: Language models usually use left-to-right (L2R) autoregressive factorization. However, L2R factorization may not always be the best inductive bias. Therefore, we investigate whether alternative factorizations of the text distribution could be beneficial in some tasks. We investigate right-to-left (R2L) training as a compelling alternative, focusing on multiple-choice questions (MCQs) as a test bed for knowledge extraction and reasoning. Through extensive experiments across various model sizes (2B-8B parameters) and training datasets, we find that R2L models can significantly outperform L2R models on several MCQ benchmarks, including logical reasoning, commonsense understanding, and truthfulness assessment tasks. Our analysis reveals that this performance difference may be fundamentally linked to multiple factors including calibration, computability and directional conditional entropy. We ablate the impact of these factors through controlled simulation studies using arithmetic tasks, where the impacting factors can be better disentangled. Our work demonstrates that exploring alternative factorizations of the text distribution can lead to improvements in LLM capabilities and provides theoretical insights into optimal factorization towards approximating human language distribution, and when each reasoning order might be more advantageous.

arXiv ID: 2502.18435v1

URL: Reversal Blessing: Thinking Backward May Outpace Thinking Forward in Multi-choice Questions

---

Paper 9: TextGames: Learning to Self-Play Text-Based Puzzle Games via Language Model Reasoning

Authors: Frederikus Hudi, Genta Indra Winata, Ruochen Zhang, Alham Fikri Aji

Published: 2025-02-25

Summary: Reasoning is a fundamental capability of large language models (LLMs), enabling them to comprehend, analyze, and solve complex problems. In this paper, we introduce TextGames, an innovative benchmark specifically crafted to assess LLMs through demanding text-based games that require advanced skills in pattern recognition, spatial awareness, arithmetic, and logical reasoning. Our analysis probes LLMs' performance in both single-turn and multi-turn reasoning, and their abilities in leveraging feedback to correct subsequent answers through self-reflection. Our findings reveal that, although LLMs exhibit proficiency in addressing most easy and medium-level problems, they face significant challenges with more difficult tasks. In contrast, humans are capable of solving all tasks when given sufficient time. Moreover, we observe that LLMs show improved performance in multi-turn predictions through self-reflection, yet they still struggle with sequencing, counting, and following complex rules consistently. Additionally, models optimized for reasoning outperform pre-trained LLMs that prioritize instruction following, highlighting the crucial role of reasoning skills in addressing highly complex problems.

arXiv ID: 2502.18431v1

URL: TextGames: Learning to Self-Play Text-Based Puzzle Games via Language Model Reasoning

---

Paper 10: PyEvalAI: AI-assisted evaluation of Jupyter Notebooks for immediate personalized feedback

Authors: Nils Wandel, David Stotko, Alexander Schier, Reinhard Klein

Published: 2025-02-25

Summary: Grading student assignments in STEM courses is a laborious and repetitive task for tutors, often requiring a week to assess an entire class. For students, this delay of feedback prevents iterating on incorrect solutions, hampers learning, and increases stress when exercise scores determine admission to the final exam. Recent advances in AI-assisted education, such as automated grading and tutoring systems, aim to address these challenges by providing immediate feedback and reducing grading workload. However, existing solutions often fall short due to privacy concerns, reliance on proprietary closed-source models, lack of support for combining Markdown, LaTeX and Python code, or excluding course tutors from the grading process. To overcome these limitations, we introduce PyEvalAI, an AI-assisted evaluation system, which automatically scores Jupyter notebooks using a combination of unit tests and a locally hosted language model to preserve privacy. Our approach is free, open-source, and ensures tutors maintain full control over the grading process. A case study demonstrates its effectiveness in improving feedback speed and grading efficiency for exercises in a university-level course on numerics.

arXiv ID: 2502.18425v1

URL: PyEvalAI: AI-assisted evaluation of Jupyter Notebooks for immediate personalized feedback

---

Paper 11: GLEAN: Generalized Category Discovery with Diverse and Quality-Enhanced LLM Feedback

Authors: Henry Peng Zou, Siffi Singh, Yi Nian, Jianfeng He, Jason Cai, Saab Mansour, Hang Su

Published: 2025-02-25

Summary: Generalized Category Discovery (GCD) is a practical and challenging open-world task that aims to recognize both known and novel categories in unlabeled data using limited labeled data from known categories. Due to the lack of supervision, previous GCD methods face significant challenges, such as difficulty in rectifying errors for confusing instances, and inability to effectively uncover and leverage the semantic meanings of discovered clusters. Therefore, additional annotations are usually required for real-world applicability. However, human annotation is extremely costly and inefficient. To address these issues, we propose GLEAN, a unified framework for generalized category discovery that actively learns from diverse and quality-enhanced LLM feedback. Our approach leverages three different types of LLM feedback to: (1) improve instance-level contrastive features, (2) generate category descriptions, and (3) align uncertain instances with LLM-selected category descriptions. Extensive experiments demonstrate the superior performance of \MethodName over state-of-the-art models across diverse datasets, metrics, and supervision settings. Our code is available at https://github.com/amazon-science/Glean.

arXiv ID: 2502.18414v1

URL: GLEAN: Generalized Category Discovery with Diverse and Quality-Enhanced LLM Feedback

---

Paper 12: When Benchmarks Talk: Re-Evaluating Code LLMs with Interactive Feedback

Authors: Jane Pan, Ryan Shar, Jacob Pfau, Ameet Talwalkar, He He, Valerie Chen

Published: 2025-02-25

Summary: Programming is a fundamentally interactive process, yet coding assistants are often evaluated using static benchmarks that fail to measure how well models collaborate with users. We introduce an interactive evaluation pipeline to examine how LLMs incorporate different types of feedback in a collaborative setting. Specifically, we perturb static coding benchmarks so that the code model must interact with a simulated user to retrieve key information about the problem. We find that interaction significantly affects model performance, as the relative rankings of 10 models across 3 datasets often vary between static and interactive settings, despite models being fairly robust to feedback that contains errors. We also observe that even when different feedback types are equally effective with respect to performance, they can impact model behaviors such as (1) how models respond to higher- vs. lower-quality feedback and (2) whether models prioritize aesthetic vs. functional edits. Our work aims to "re-evaluate" model coding capabilities through an interactive lens toward bridging the gap between existing evaluations and real-world usage.

arXiv ID: 2502.18413v1

URL: When Benchmarks Talk: Re-Evaluating Code LLMs with Interactive Feedback

---

Paper 13: TSKANMixer: Kolmogorov-Arnold Networks with MLP-Mixer Model for Time Series Forecasting

Authors: Young-Chae Hong, Bei Xiao, Yangho Chen

Published: 2025-02-25

Summary: Time series forecasting has long been a focus of research across diverse fields, including economics, energy, healthcare, and traffic management. Recent works have introduced innovative architectures for time series models, such as the Time-Series Mixer (TSMixer), which leverages multi-layer perceptrons (MLPs) to enhance prediction accuracy by effectively capturing both spatial and temporal dependencies within the data. In this paper, we investigate the capabilities of the Kolmogorov-Arnold Networks (KANs) for time-series forecasting by modifying TSMixer with a KAN layer (TSKANMixer). Experimental results demonstrate that TSKANMixer tends to improve prediction accuracy over the original TSMixer across multiple datasets, ranking among the top-performing models compared to other time series approaches. Our results show that the KANs are promising alternatives to improve the performance of time series forecasting by replacing or extending traditional MLPs.

arXiv ID: 2502.18410v1

URL: TSKANMixer: Kolmogorov-Arnold Networks with MLP-Mixer Model for Time Series Forecasting

---

Paper 14: AgentRM: Enhancing Agent Generalization with Reward Modeling

Authors: Yu Xia, Jingru Fan, Weize Chen, Siyu Yan, Xin Cong, Zhong Zhang, Yaxi Lu, Yankai Lin, Zhiyuan Liu, Maosong Sun

Published: 2025-02-25

Summary: Existing LLM-based agents have achieved strong performance on held-in tasks, but their generalizability to unseen tasks remains poor. Hence, some recent work focus on fine-tuning the policy model with more diverse tasks to improve the generalizability. In this work, we find that finetuning a reward model to guide the policy model is more robust than directly finetuning the policy model. Based on this finding, we propose AgentRM, a generalizable reward model, to guide the policy model for effective test-time search. We comprehensively investigate three approaches to construct the reward model, including explicit reward modeling, implicit reward modeling and LLM-as-a-judge. We then use AgentRM to guide the answer generation with Best-of-N sampling and step-level beam search. On four types of nine agent tasks, AgentRM enhances the base policy model by 8.8 points on average, surpassing the top general agent by 4.0. Moreover, it demonstrates weak-to-strong generalization, yielding greater improvement of 12.6 on LLaMA-3-70B policy model. As for the specializability, AgentRM can also boost a finetuned policy model and outperform the top specialized agent by 11.4 on three held-in tasks. Further analysis verifies its effectiveness in test-time scaling. Codes will be released to facilitate the research in this area.

arXiv ID: 2502.18407v1

URL: AgentRM: Enhancing Agent Generalization with Reward Modeling

---

Paper 15: The Gradient of Algebraic Model Counting

Authors: Jaron Maene, Luc De Raedt

Published: 2025-02-25

Summary: Algebraic model counting unifies many inference tasks on logic formulas by exploiting semirings. Rather than focusing on inference, we consider learning, especially in statistical-relational and neurosymbolic AI, which combine logical, probabilistic and neural representations. Concretely, we show that the very same semiring perspective of algebraic model counting also applies to learning. This allows us to unify various learning algorithms by generalizing gradients and backpropagation to different semirings. Furthermore, we show how cancellation and ordering properties of a semiring can be exploited for more memory-efficient backpropagation. This allows us to obtain some interesting variations of state-of-the-art gradient-based optimisation methods for probabilistic logical models. We also discuss why algebraic model counting on tractable circuits does not lead to more efficient second-order optimization. Empirically, our algebraic backpropagation exhibits considerable speed-ups as compared to existing approaches.

arXiv ID: 2502.18406v1

URL: The Gradient of Algebraic Model Counting

Paper 16: "Why do we do this?": Moral Stress and the Affective Experience of Ethics in Practice

Authors: Sonja Rattay, Ville Vakkuri, Marco Rozendaal, Irina Shklovski

Published: 2025-02-25

Summary: A plethora of toolkits, checklists, and workshops have been developed to bridge the well-documented gap between AI ethics principles and practice. Yet little is known about effects of such interventions on practitioners. We conducted an ethnographic investigation in a major European city organization that developed and works to integrate an ethics toolkit into city operations. We find that the integration of ethics tools by technical teams destabilises their boundaries, roles, and mandates around responsibilities and decisions. This lead to emotional discomfort and feelings of vulnerability, which neither toolkit designers nor the organization had accounted for. We leverage the concept of moral stress to argue that this affective experience is a core challenge to the successful integration of ethics tools in technical practice. Even in this best case scenario, organisational structures were not able to deal with moral stress that resulted from attempts to implement responsible technology development practices.

arXiv ID: 2502.18395v1

URL: "Why do we do this?": Moral Stress and the Affective Experience of Ethics in Practice

---

Paper 17: Monte Carlo Temperature: a robust sampling strategy for LLM's uncertainty quantification methods

Authors: Nicola Cecere, Andrea Bacciu, Ignacio Fernández Tobías, Amin Mantrach

Published: 2025-02-25

Summary: Uncertainty quantification (UQ) in Large Language Models (LLMs) is essential for their safe and reliable deployment, particularly in critical applications where incorrect outputs can have serious consequences. Current UQ methods typically rely on querying the model multiple times using non-zero temperature sampling to generate diverse outputs for uncertainty estimation. However, the impact of selecting a given temperature parameter is understudied, and our analysis reveals that temperature plays a fundamental role in the quality of uncertainty estimates. The conventional approach of identifying optimal temperature values requires expensive hyperparameter optimization (HPO) that must be repeated for each new model-dataset combination. We propose Monte Carlo Temperature (MCT), a robust sampling strategy that eliminates the need for temperature calibration. Our analysis reveals that: 1) MCT provides more robust uncertainty estimates across a wide range of temperatures, 2) MCT improves the performance of UQ methods by replacing fixed-temperature strategies that do not rely on HPO, and 3) MCT achieves statistical parity with oracle temperatures, which represent the ideal outcome of a well-tuned but computationally expensive HPO process. These findings demonstrate that effective UQ can be achieved without the computational burden of temperature parameter calibration.

arXiv ID: 2502.18389v1

URL: Monte Carlo Temperature: a robust sampling strategy for LLM's uncertainty quantification methods

---

Paper 18: How Far are LLMs from Real Search? A Comprehensive Study on Efficiency, Completeness, and Inherent Capabilities

Authors: Minhua Lin, Hui Liu, Xianfeng Tang, Jingying Zeng, Zhenwei Dai, Chen Luo, Zheng Li, Xiang Zhang, Qi He, Suhang Wang

Published: 2025-02-25

Summary: Search plays a fundamental role in problem-solving across various domains, with most real-world decision-making problems being solvable through systematic search. Drawing inspiration from recent discussions on search and learning, we systematically explore the complementary relationship between search and Large Language Models (LLMs) from three perspectives. First, we analyze how learning can enhance search efficiency and propose Search via Learning (SeaL), a framework that leverages LLMs for effective and efficient search. Second, we further extend SeaL to SeaL-C to ensure rigorous completeness during search. Our evaluation across three real-world planning tasks demonstrates that SeaL achieves near-perfect accuracy while reducing search spaces by up to 99.1% compared to traditional approaches. Finally, we explore how far LLMs are from real search by investigating whether they can develop search capabilities independently. Our analysis reveals that while current LLMs struggle with efficient search in complex problems, incorporating systematic search strategies significantly enhances their problem-solving capabilities. These findings not only validate the effectiveness of our approach but also highlight the need for improving LLMs' search abilities for real-world applications.

arXiv ID: 2502.18387v1

URL: How Far are LLMs from Real Search? A Comprehensive Study on Efficiency, Completeness, and Inherent Capabilities

---

Paper 19: Semantic and Goal-oriented Wireless Network Coverage: The Area of Effectiveness

Authors: Mattia Merluzzi, Giuseppe Di Poce, Paolo Di Lorenzo

Published: 2025-02-25

Summary: Assessing wireless coverage is a fundamental task for public network operators and private deployments, whose goal is to guarantee quality of service across the network while minimizing material waste and energy consumption. These maps are usually built through ray tracing techniques and/or channel measurements that can be consequently translated into network Key Performance Indicators (KPIs), such as capacity or throughput. However, next generation networks (e.g., 6G) typically involve beyond communication resources, towards services that require data transmission, but also processing (local and remote) to perform complex decision making in real time, with the best balance between performance, energy consumption, material waste, and privacy. In this paper, we introduce the novel concept of areas of effectiveness, which goes beyond the legacy notion of coverage, towards one that takes into account capability of the network of offering edge Artificial Intelligence (AI)-related computation. We will show that radio coverage is a poor indicator of real system performance, depending on the application and the computing capabilities of network and devices. This opens new challenges in network planning, but also resource orchestration during operation to achieve the specific goal of communication.

arXiv ID: 2502.18381v1

URL: Semantic and Goal-oriented Wireless Network Coverage: The Area of Effectiveness

---

Paper 20: MindMem: Multimodal for Predicting Advertisement Memorability Using LLMs and Deep Learning

Authors: Sepehr Asgarian, Qayam Jetha, Jouhyun Jeon

Published: 2025-02-25

Summary: In the competitive landscape of advertising, success hinges on effectively navigating and leveraging complex interactions among consumers, advertisers, and advertisement platforms. These multifaceted interactions compel advertisers to optimize strategies for modeling consumer behavior, enhancing brand recall, and tailoring advertisement content. To address these challenges, we present MindMem, a multimodal predictive model for advertisement memorability. By integrating textual, visual, and auditory data, MindMem achieves state-of-the-art performance, with a Spearman's correlation coefficient of 0.631 on the LAMBDA and 0.731 on the Memento10K dataset, consistently surpassing existing methods. Furthermore, our analysis identified key factors influencing advertisement memorability, such as video pacing, scene complexity, and emotional resonance. Expanding on this, we introduced MindMem-ReAd (MindMem-Driven Re-generated Advertisement), which employs Large Language Model-based simulations to optimize advertisement content and placement, resulting in up to a 74.12% improvement in advertisement memorability. Our results highlight the transformative potential of Artificial Intelligence in advertising, offering advertisers a robust tool to drive engagement, enhance competitiveness, and maximize impact in a rapidly evolving market.

arXiv ID: 2502.18371v1

URL: MindMem: Multimodal for Predicting Advertisement Memorability Using LLMs and Deep Learning

---

Summary of Findings

This section would contain a concise summary of the research findings across all reviewed papers. It would discuss key trends, challenges, and opportunities related to using LoRA on smartphones. For instance, it might highlight the effectiveness of different LoRA implementations for specific tasks, the impact of model size and quantization on performance, and future research directions. Specific papers would be cited using the IEEE style.