Paper Watch (arXiv)
crosslingual-factual-consistency

Leo Labat

2025-05-06

1 Measuring Hong Kong Massive Multi-Task Language Understanding

Chuxue Cao, Zhenghao Zhu, Junqi Zhu, Guoying Lu, Siyu Peng, Juntao Dai, Weijie Shi, Sirui Han, Yike Guo
https://arxiv.org/abs/arXiv:2505.02177
Multilingual understanding is crucial for the cross-cultural applicability of Large Language Models (LLMs). However, evaluation benchmarks designed for Hong Kong's unique linguistic landscape, which combines Traditional Chinese script with Cantonese as the spoken form and its cultural context, remain underdeveloped. To address this gap, we introduce HKMMLU, a multi-task language understanding benchmark that evaluates Hong Kong's linguistic competence and socio-cultural knowledge. The HKMMLU includes 26,698 multi-choice questions across 66 subjects, organized into four categories: Science, Technology, Engineering, and Mathematics (STEM), Social Sciences, Humanities, and Other. To evaluate the multilingual understanding ability of LLMs, 90,550 Mandarin-Cantonese translation tasks were additionally included. We conduct comprehensive experiments on GPT-4o, Claude 3.7 Sonnet, and 18 open-source LLMs of varying sizes on HKMMLU. The results show that the best-performing model, DeepSeek-V3, struggles to achieve an accuracy of 75\%, significantly lower than that of MMLU and CMMLU. This performance gap highlights the need to improve LLMs' capabilities in Hong Kong-specific language and knowledge domains. Furthermore, we investigate how question language, model size, prompting strategies, and question and reasoning token lengths affect model performance. We anticipate that HKMMLU will significantly advance the development of LLMs in multilingual and cross-cultural contexts, thereby enabling broader and more impactful applications.

2 Large Language Models Understanding: an Inherent Ambiguity Barrier

Daniel N. Nissani (Nissensohn)
https://arxiv.org/abs/arXiv:2505.00654
A lively ongoing debate is taking place, since the extraordinary emergence of Large Language Models (LLMs) with regards to their capability to understand the world and capture the meaning of the dialogues in which they are involved. Arguments and counter-arguments have been proposed based upon thought experiments, anecdotal

conversations between LLMs and humans, statistical linguistic analysis, philosophical considerations, and more. In this brief paper we present a counter-argument based upon a thought experiment and semi-formal considerations leading to an inherent ambiguity barrier which prevents LLMs from having any understanding of what their amazingly fluent dialogues mean.

3 Inducing Robustness in a 2 Dimensional Direct Preference Optimization Paradigm

Sarvesh Shashidhar, Ritik, Nachiketa Patil, Suraj Racha, Ganesh Ramakrishnan

https://arxiv.org/abs/arXiv:2505.01706

Direct Preference Optimisation (DPO) has emerged as a powerful method for aligning Large Language Models (LLMs) with human preferences, offering a stable and efficient alternative to approaches that use Reinforcement learning via Human Feedback. In this work, we investigate the performance of DPO using open-source preference datasets. One of the major drawbacks of DPO is that it doesn't induce granular scoring and treats all the segments of the responses with equal propensity. However, this is not practically true for human preferences since even "good" responses have segments that may not be preferred by the annotator. To resolve this, a 2-dimensional scoring for DPO alignment called 2D-DPO was proposed. We explore the 2D-DPO alignment paradigm and the advantages it provides over the standard DPO by comparing their win rates. It is observed that these methods, even though effective, are not robust to label/score noise. To counter this, we propose an approach of incorporating segment-level score noise robustness to the 2D-DPO algorithm. Along with theoretical backing, we also provide empirical verification in favour of the algorithm and introduce other noise models that can be present.

4 Always Tell Me The Odds: Fine-grained Conditional Probability Estimation

Liaoyaqi Wang, Zhengping Jiang, Anqi Liu, Benjamin Van Durme

https://arxiv.org/abs/arXiv:2505.01595

We present a state-of-the-art model for fine-grained probability estimation of propositions conditioned on context. Recent advances in large language models (LLMs) have significantly enhanced their reasoning capabilities, particularly on well-defined tasks with complete information. However, LLMs continue to struggle with making accurate and well-calibrated probabilistic predictions under uncertainty or partial information. While incorporating uncertainty into model predictions often boosts performance, obtaining reliable estimates of that uncertainty remains understudied. In particular, LLM probability estimates tend to be coarse and biased towards more frequent numbers. Through a combination of human and synthetic data creation and assessment, scaling to larger models, and better supervision, we propose a set of strong and precise prob-

ability estimation models. We conduct systematic evaluations across tasks that rely on conditional probability estimation and show that our approach consistently outperforms existing fine-tuned and prompting-based methods by a large margin.

5 $\textit{New News}$: System-2 Fine-tuning for Robust Integration of New Knowledge

Core Francisco Park, Zechen Zhang, Hidenori Tanaka
https://arxiv.org/abs/arXiv:2505.01812
Humans and intelligent animals can effortlessly internalize new information ("news") and accurately extract the implications for performing downstream tasks. While large language models (LLMs) can achieve this through in-context learning (ICL) when the news is explicitly given as context, fine-tuning remains challenging for the models to consolidate learning in weights. In this paper, we introduce $\textit{New News}$, a dataset composed of hypothetical yet plausible news spanning multiple domains (mathematics, coding, discoveries, leaderboards, events), accompanied by downstream evaluation questions whose correct answers critically depend on understanding and internalizing the news. We first demonstrate a substantial gap between naive fine-tuning and in-context learning (FT-ICL gap) on our news dataset. To address this gap, we explore a suite of self-play data generation protocols -- paraphrases, implications and Self-QAs -- designed to distill the knowledge from the model with context into the weights of the model without the context, which we term $\textit{System-2 Fine-tuning}$ (Sys2-FT). We systematically evaluate ICL and Sys2-FT performance across data domains and model scales with the Qwen 2.5 family of models. Our results demonstrate that the self-QA protocol of Sys2-FT significantly improves models' in-weight learning of the news. Furthermore, we discover the $\textit{contexual shadowing effect}$, where training with the news $\textit{in context}$ followed by its rephrases or QAs degrade learning of the news. Finally, we show preliminary evidence of an emerging scaling law of Sys2-FT.

6 RM-R1: Reward Modeling as Reasoning

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, Heng Ji
https://arxiv.org/abs/arXiv:2505.02387
Reward modeling is essential for aligning large language models (LLMs) with human preferences, especially through reinforcement learning from human feedback (RLHF). To provide accurate reward signals, a reward model (RM) should stimulate deep thinking and conduct interpretable reasoning before assigning a score or a judgment. However, existing RMs either produce opaque scalar scores or directly generate the prediction of a preferred answer, making them struggle to integrate natural language critiques, thus lacking interpretability. Inspired by recent advances of long chain-of-thought

(CoT) on reasoning-intensive tasks, we hypothesize and validate that integrating reasoning capabilities into reward modeling significantly enhances RM's interpretability and performance. In this work, we introduce a new class of generative reward models -- Reasoning Reward Models (ReasRMs) -- which formulate reward modeling as a reasoning task. We propose a reasoning-oriented training pipeline and train a family of ReasRMs, RM-R1. The training consists of two key stages: (1) distillation of high-quality reasoning chains and (2) reinforcement learning with verifiable rewards. RM-R1 improves LLM rollouts by self-generating reasoning traces or chat-specific rubrics and evaluating candidate responses against them. Empirically, our models achieve state-of-the-art or near state-of-the-art performance of generative RMs across multiple comprehensive reward model benchmarks, outperforming much larger open-weight models (e.g., Llama3.1-405B) and proprietary ones (e.g., GPT-4o) by up to 13.8%. Beyond final performance, we perform thorough empirical analysis to understand the key ingredients of successful ReasRM training. To facilitate future research, we release six ReasRM models along with code and data at this https URL.

7 Rosetta-PL: Propositional Logic as a Benchmark for Large Language Model Reasoning

Shaun Baek, Shaun Esua-Mensah, Cyrus Tsui, Sejan Vigneswaralingam, Abdullah Alali, Michael Lu, Vasu Sharma, Sean O'Brien, Kevin Zhu
https://arxiv.org/abs/arXiv:2505.00001
Large Language Models (LLMs) are primarily trained on high-resource natural languages, limiting their effectiveness in low-resource settings and in tasks requiring deep logical reasoning. This research introduces Rosetta-PL, a benchmark designed to evaluate LLMs' logical reasoning and generalization capabilities in a controlled environment. We construct Rosetta-PL by translating a dataset of logical propositions from Lean into a custom logical language, which is then used to fine-tune an LLM (e.g., GPT-4o). Our experiments analyze the impact of the size of the dataset and the translation methodology on the performance of the model. Our results indicate that preserving logical relationships in the translation process significantly boosts precision, with accuracy plateauing beyond roughly 20,000 training samples. These insights provide valuable guidelines for optimizing LLM training in formal reasoning tasks and improving performance in various low-resource language applications.

8 SIMPLEMIX: Frustratingly Simple Mixing of Off- and On-policy Data in Language Model Preference Learning

Tianjian Li, Daniel Khashabi
https://arxiv.org/abs/arXiv:2505.02363
Aligning language models with human preferences relies on pairwise preference datasets. While some studies suggest that on-policy data consistently outperforms off -policy data for preference learning, others indicate that the advantages of on-policy data may be

task-dependent, highlighting the need for a systematic exploration of their interplay. In this work, we show that on-policy and off-policy data offer complementary strengths in preference optimization: on-policy data is particularly effective for reasoning tasks like math and coding, while off-policy data performs better on open-ended tasks such as creative writing and making personal recommendations. Guided by these findings, we introduce SIMPLEMIX, an approach to combine the complementary strengths of on-policy and off-policy preference learning by simply mixing these two data sources. Our empirical results across diverse tasks and benchmarks demonstrate that SIMPLEMIX substantially improves language model alignment. Specifically, SIMPLEMIX improves upon on-policy DPO and off-policy DPO by an average of 6.03% on Alpaca Eval 2.0. Moreover, it outperforms prior approaches that are much more complex in combining on- and off-policy data, such as HyPO and DPO-Mix-P, by an average of 3.05%.

9 Semantic Volume: Quantifying and Detecting both External and Internal Uncertainty in LLMs

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, Anurag Beniwal

https://arxiv.org/abs/arXiv:2502.21239

Large language models (LLMs) have demonstrated remarkable performance across diverse tasks by encoding vast amounts of factual knowledge. However, they are still prone to hallucinations, generating incorrect or misleading information, often accompanied by high uncertainty. Existing methods for hallucination detection primarily focus on quantifying internal uncertainty, which arises from missing or conflicting knowledge within the model. However, hallucinations can also stem from external uncertainty, where ambiguous user queries lead to multiple possible interpretations. In this work, we introduce Semantic Volume, a novel mathematical measure for quantifying both external and internal uncertainty in LLMs. Our approach perturbs queries and responses, embeds them in a semantic space, and computes the determinant of the Gram matrix of the embedding vectors, capturing their dispersion as a measure of uncertainty. Our framework provides a generalizable and unsupervised uncertainty detection method without requiring internal access to LLMs. We conduct extensive experiments on both external and internal uncertainty detection, demonstrating that our Semantic Volume method consistently outperforms existing baselines in both tasks. Additionally, we provide theoretical insights linking our measure to differential entropy, unifying and extending previous sampling-based uncertainty measures such as the semantic entropy. Semantic Volume is shown to be a robust and interpretable approach to improving the reliability of LLMs by systematically detecting uncertainty in both user queries and model responses.

10 What do Language Model Probabilities Represent? From Distribution Estimation to Response Prediction

Eitan Wagner, Omri Abend
https://arxiv.org/abs/arXiv:2505.02072
The notion of language modeling has gradually shifted in recent years from a distribution over finite-length strings to general-purpose prediction models for textual inputs and outputs, following appropriate alignment phases. This paper analyzes the distinction between distribution estimation and response prediction in the context of LLMs, and their often conflicting goals. We examine the training phases of LLMs, which include pretraining, in-context learning, and preference tuning, and also the common use cases for their output probabilities, which include completion probabilities and explicit probabilities as output. We argue that the different settings lead to three distinct intended output distributions. We demonstrate that NLP works often assume that these distributions should be similar, which leads to misinterpretations of their experimental findings. Our work sets firmer formal foundations for the interpretation of LLMs, which will inform ongoing work on the interpretation and use of LLMs' induced distributions.

11 Invoke Interfaces Only When Needed: Adaptive Invocation for Large Language Models in Question Answering

Jihao Zhao, Chunlai Zhou, Biao Qin
https://arxiv.org/abs/arXiv:2505.02311
The collaborative paradigm of large and small language models (LMs) effectively balances performance and cost, yet its pivotal challenge lies in precisely pinpointing the moment of invocation when hallucinations arise in small LMs. Previous optimization efforts primarily focused on post-processing techniques, which were separate from the reasoning process of LMs, resulting in high computational costs and limited effectiveness. In this paper, we propose a practical invocation evaluation metric called AttenHScore, which calculates the accumulation and propagation of hallucinations during the generation process of small LMs, continuously amplifying potential reasoning errors. By dynamically adjusting the detection threshold, we achieve more accurate real-time invocation of large LMs. Additionally, considering the limited reasoning capacity of small LMs, we leverage uncertainty-aware knowledge reorganization to assist them better capture critical information from different text chunks. Extensive experiments reveal that our AttenHScore outperforms most baseline in enhancing real-time hallucination detection capabilities across multiple QA datasets, especially when addressing complex queries. Moreover, our strategies eliminate the need for additional model training and display flexibility in adapting to various transformer-based LMs.

12 Demystifying optimized prompts in language models

Rimon Melamed, Lucas H. McCabe, H. Howie Huang
https://arxiv.org/abs/arXiv:2505.02273
Modern language models (LMs) are not robust to out-of-distribution inputs. Machine generated ("optimized") prompts can be used to modulate LM outputs and induce specific behaviors while appearing completely uninterpretable. In this work, we investigate the composition of optimized prompts, as well as the mechanisms by which LMs parse and build predictions from optimized prompts. We find that optimized prompts primarily consist of punctuation and noun tokens which are more rare in the training data. Internally, optimized prompts are clearly distinguishable from natural language counterparts based on sparse subsets of the model's activations. Across various families of instruction-tuned models, optimized prompts follow a similar path in how their representations form through the network.

13 LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, Ziwei Liu
https://arxiv.org/abs/arXiv:2407.12772
The advances of large foundation models necessitate wide-coverage, low-cost, and zero-contamination benchmarks. Despite continuous exploration of language model evaluations, comprehensive studies on the evaluation of Large Multi-modal Models (LMMs) remain limited. In this work, we introduce LMMS-EVAL, a unified and standardized multimodal benchmark framework with over 50 tasks and more than 10 models to promote transparent and reproducible evaluations. Although LMMS-EVAL offers comprehensive coverage, we find it still falls short in achieving low cost and zero contamination. To approach this evaluation trilemma, we further introduce LMMS-EVAL LITE, a pruned evaluation toolkit that emphasizes both coverage and efficiency. Additionally, we present Multimodal LIVEBENCH that utilizes continuously updating news and online forums to assess models' generalization abilities in the wild, featuring a low-cost and zero-contamination evaluation approach. In summary, our work highlights the importance of considering the evaluation trilemma and provides practical solutions to navigate the trade-offs in evaluating large multi-modal models, paving the way for more effective and reliable benchmarking of LMMs. We opensource our codebase and maintain leaderboard of LIVEBENCH at this https URL and this https URL.

14 ELOQ: Resources for Enhancing LLM Detection of Out-of-Scope Questions

Zhiyuan Peng, Jinming Nian, Alexandre Evfimievski, Yi Fang
https://arxiv.org/abs/arXiv:2410.14567
Retrieval-augmented generation (RAG) has become integral to large language models (LLMs), particularly for conversational AI systems where user questions may reference knowledge beyond the LLMs' training cutoff. However, many natural user questions lack well-defined answers, either due to limited domain knowledge or because the retrieval system returns documents that are relevant in appearance but uninformative in content. In such cases, LLMs often produce hallucinated answers without flagging them. While recent work has largely focused on questions with false premises, we study out-of-scope questions, where the retrieved document appears semantically similar to the question but lacks the necessary information to answer it. In this paper, we propose a guided hallucination-based approach ELOQ to automatically generate a diverse set of out-of-scope questions from post-cutoff documents, followed by human verification to ensure quality. We use this dataset to evaluate several LLMs on their ability to detect out-of-scope questions and generate appropriate responses. Finally, we introduce an improved detection method that enhances the reliability of LLM-based question-answering systems in handling out-of-scope questions.

15 LLMs for Extremely Low-Resource Finno-Ugric Languages

Taido Purason, Hele-Andra Kuulmets, Mark Fishel
https://arxiv.org/abs/arXiv:2410.18902
The advancement of large language models (LLMs) has predominantly focused on high-resource languages, leaving low-resource languages, such as those in the Finno-Ugric family, significantly underrepresented. This paper addresses this gap by focusing on Võro, Livonian, and Komi. We cover almost the entire cycle of LLM creation, from data collection to instruction tuning and evaluation. Our contributions include developing multilingual base and instruction-tuned models; creating evaluation benchmarks, including the smugri-MT-bench multi-turn conversational benchmark; and conducting human evaluation. We intend for this work to promote linguistic diversity, ensuring that lesser-resourced languages can benefit from advancements in NLP.

16 Optimizing Chain-of-Thought Reasoners via Gradient Variance Minimization in Rejection Sampling and RL

Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, Tong Zhang
https://arxiv.org/abs/arXiv:2505.02391
Chain-of-thought (CoT) reasoning in large language models (LLMs) can be formalized as a latent variable problem, where the model needs to generate intermediate reason-

ing steps. While prior approaches such as iterative reward-ranked fine-tuning (RAFT) have relied on such formulations, they typically apply uniform inference budgets across prompts, which fails to account for variability in difficulty and convergence behavior. This work identifies the main bottleneck in CoT training as inefficient stochastic gradient estimation due to static sampling strategies. We propose GVM-RAFT, a prompt-specific Dynamic Sample Allocation Strategy designed to minimize stochastic gradient variance under a computational budget constraint. The method dynamically allocates computational resources by monitoring prompt acceptance rates and stochastic gradient norms, ensuring that the resulting gradient variance is minimized. Our theoretical analysis shows that the proposed dynamic sampling strategy leads to accelerated convergence guarantees under suitable conditions. Experiments on mathematical reasoning show that GVM-RAFT achieves a 2-4x speedup and considerable accuracy improvements over vanilla RAFT. The proposed dynamic sampling strategy is general and can be incorporated into other reinforcement learning algorithms, such as GRPO, leading to similar improvements in convergence and test accuracy. Our code is available at this https URL.

17 Unveiling the Mechanisms of Explicit CoT Training: How CoT Enhances Reasoning Generalization

Xinhao Yao, Ruifeng Ren, Yun Liao, Yong Liu

https://arxiv.org/abs/arXiv:2502.04667

The integration of explicit Chain-of-Thought (CoT) reasoning into training large language models (LLMs) has advanced their reasoning capabilities, yet the mechanisms by which CoT enhances generalization remain poorly understood. This work investigates (1) \textit{how} CoT training reshapes internal model representations and (2) \textit{why} it improves both in-distribution (ID) and out-of-distribution (OOD) reasoning generalization. Through controlled experiments and theoretical analysis, we derive the following key insights. \textbf{(1)} Structural Advantage: CoT training internalizes reasoning into a two-stage generalizing circuit, where the number of stages corresponds to the explicit reasoning steps during training. Notably, CoT-trained models resolve intermediate results at shallower layers compared to non-CoT counterparts, freeing up deeper layers to specialize in subsequent reasoning steps. \textbf{(2)} Theoretical Analysis: the information-theoretic generalization bounds via distributional divergence can be decomposed into ID and OOD components. While ID error diminishes with sufficient training regardless of CoT, OOD error critically depends on CoT: Non-CoT training fails to generalize to OOD samples due to unseen reasoning patterns, whereas CoT training achieves near-perfect OOD generalization by mastering subtasks and reasoning compositions during training. The identified mechanisms explain our experimental results: CoT training accelerates convergence and enhances generalization from ID to both ID and OOD scenarios while maintaining robust performance even with tolerable noise. These findings are further validated on complex real-world datasets. This paper offers valuable insights for designing CoT strategies to enhance LLM reasoning robustness.

18 Personalisation or Prejudice? Addressing Geographic Bias in Hate Speech Detection using Debias Tuning in Large Language Models

Paloma Piot, Patricia Martín-Rodilla, Javier Parapar
https://arxiv.org/abs/arXiv:2505.02252
Commercial Large Language Models (LLMs) have recently incorporated memory features to deliver personalised responses. This memory retains details such as user demographics and individual characteristics, allowing LLMs to adjust their behaviour based on personal information. However, the impact of integrating personalised information into the context has not been thoroughly assessed, leading to questions about its influence on LLM behaviour. Personalisation can be challenging, particularly with sensitive topics. In this paper, we examine various state-of-the-art LLMs to understand their behaviour in different personalisation scenarios, specifically focusing on hate speech. We prompt the models to assume country-specific personas and use different languages for hate speech detection. Our findings reveal that context personalisation significantly influences LLMs' responses in this sensitive area. To mitigate these unwanted biases, we fine-tune the LLMs by penalising inconsistent hate speech classifications made with and without country or language-specific context. The refined models demonstrate improved performance in both personalised contexts and when no context is provided.

19 A Survey on Test-Time Scaling in Large Language Models: What, How, Where, and How Well?

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, Chen Ma
https://arxiv.org/abs/arXiv:2503.24235
As enthusiasm for scaling computation (data and parameters) in the pretraining era gradually diminished, test-time scaling (TTS), also referred to as "test-time computing" has emerged as a prominent research focus. Recent studies demonstrate that TTS can further elicit the problem-solving capabilities of large language models (LLMs), enabling significant breakthroughs not only in specialized reasoning tasks, such as mathematics and coding, but also in general tasks like open-ended Q&A. However, despite the explosion of recent efforts in this area, there remains an urgent need for a comprehensive survey offering a systemic understanding. To fill this gap, we propose a unified, multidimensional framework structured along four core dimensions of TTS research: what to scale, how to scale, where to scale, and how well to scale. Building upon this taxonomy, we conduct an extensive review of methods, application scenarios, and assessment aspects, and present an organized decomposition that highlights the unique functional roles of individual techniques within the broader TTS landscape. From this analysis, we distill the major developmental trajectories of TTS to date and offer hands-on guidelines for practical deployment. Furthermore, we identify several open challenges and offer insights into promising future directions, including further scaling, clarifying the functional essence of techniques, generalizing to more tasks, and more attributions.

Our repository is available on this https URL

20 Exploring the Potential of Offline RL for Reasoning in LLMs: A Preliminary Study

Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Yunjie Ji, Han Zhao, Xiangang Li

https://arxiv.org/abs/arXiv:2505.02142

Despite significant advances in long-context reasoning by large language models (LLMs), primarily through Online Reinforcement Learning (RL) methods, these approaches incur substantial computational costs and complexity. In contrast, simpler and more economical Offline RL methods remain underexplored. To address this gap, we investigate the effectiveness of Offline RL methods, specifically Direct Preference Optimization (DPO) and its length-desensitized variant LD-DPO, in enhancing the reasoning capabilities of LLMs. Extensive experiments across multiple reasoning benchmarks demonstrate that these simpler Offline RL methods substantially improve model performance, achieving an average enhancement of 3.3\%, with a particularly notable increase of 10.1\% on the challenging Arena-Hard benchmark. Furthermore, we analyze DPO's sensitivity to output length, emphasizing that increasing reasoning length should align with semantic richness, as indiscriminate lengthening may adversely affect model performance. We provide comprehensive descriptions of our data processing and training methodologies, offering empirical evidence and practical insights for developing more cost-effective Offline RL approaches.

21 Same evaluation, more tokens: On the effect of input length for machine translation evaluation using Large Language Models

Tobias Domhan, Dawei Zhu

https://arxiv.org/abs/arXiv:2505.01761

Accurately evaluating machine-translated text remains a long-standing challenge, particularly for long documents. Recent work has shown that large language models (LLMs) can serve as reliable and interpretable sentence-level translation evaluators via MQM error span annotations. With modern LLMs supporting larger context windows, a natural question arises: can we feed entire document translations into an LLM for quality assessment? Ideally, evaluation should be invariant to text length, producing consistent error spans regardless of input granularity. However, our analysis shows that text length significantly impacts evaluation: longer texts lead to fewer error spans and reduced system ranking accuracy. To address this limitation, we evaluate several strategies, including granularity-aligned prompting, Focus Sentence Prompting (FSP), and a fine-tuning approach to better align LLMs with the evaluation task. The latter two methods largely mitigate this length bias, making LLMs more reliable for long-form translation evaluation.

22 Multilingual Brain Surgeon: Large Language Models Can be Compressed Leaving No Language Behind

Hongchuan Zeng, Hongshen Xu, Lu Chen, Kai Yu
https://arxiv.org/abs/arXiv:2404.04748
Large Language Models (LLMs) have ushered in a new era in Natural Language Processing, but their massive size demands effective compression techniques for practicality. Although numerous model compression techniques have been investigated, they typically rely on a calibration set that overlooks the multilingual context and results in significant accuracy degradation for low-resource languages. This paper introduces Multilingual Brain Surgeon (MBS), a novel calibration data sampling method for multilingual LLMs compression. MBS overcomes the English-centric limitations of existing methods by sampling calibration data from various languages proportionally to the language distribution of the model training datasets. Our experiments, conducted on the BLOOM multilingual LLM, demonstrate that MBS improves the performance of existing English-centric compression methods, especially for low-resource languages. We also uncover the dynamics of language interaction during compression, revealing that the larger the proportion of a language in the training set and the more similar the language is to the calibration language, the better performance the language retains after compression. In conclusion, MBS presents an innovative approach to compressing multilingual LLMs, addressing the performance disparities and improving the language inclusivity of existing compression techniques.

23 AI agents may be worth the hype but not the resources (yet): An initial exploration of machine translation quality and costs in three language pairs in the legal and news domains

Vicent Briva Iglesias, Gokhan Dogru
https://arxiv.org/abs/arXiv:2505.01560
Large language models (LLMs) and multi-agent orchestration are touted as the next leap in machine translation (MT), but their benefits relative to conventional neural MT (NMT) remain unclear. This paper offers an empirical reality check. We benchmark five paradigms, Google Translate (strong NMT baseline), GPT-4o (general-purpose LLM), o1-preview (reasoning-enhanced LLM), and two GPT-4o-powered agentic workflows (sequential three-stage and iterative refinement), on test data drawn from a legal contract and news prose in three English-source pairs: Spanish, Catalan and Turkish. Automatic evaluation is performed with COMET, BLEU, chrF2 and TER; human evaluation is conducted with expert ratings of adequacy and fluency; efficiency with total input-plus-output token counts mapped to April 2025 pricing.
Automatic scores still favour the mature NMT system, which ranks first in seven of

twelve metric-language combinations; o1-preview ties or places second in most remaining cases, while both multi-agent workflows trail. Human evaluation reverses part of this narrative: o1-preview produces the most adequate and fluent output in five of six comparisons, and the iterative agent edges ahead once, indicating that reasoning layers capture semantic nuance undervalued by surface metrics. Yet these qualitative gains carry steep costs. The sequential agent consumes roughly five times, and the iterative agent fifteen times, the tokens used by NMT or single-pass LLMs.
We advocate multidimensional, cost-aware evaluation protocols and highlight research directions that could tip the balance: leaner coordination strategies, selective agent activation, and hybrid pipelines combining single-pass LLMs with targeted agent intervention.

24 Dynamic Parametric Retrieval Augmented Generation for Test-time Knowledge Enhancement

Yuqiao Tan, Shizhu He, Huanxuan Liao, Jun Zhao, Kang Liu
https://arxiv.org/abs/arXiv:2503.23895
Retrieval-augmented generation (RAG) enhances large language models (LLMs) by retrieving relevant documents from external sources and incorporating them into the context. While it improves reliability by providing factual texts, it significantly increases inference costs as context length grows and introduces challenging issue of RAG hallucination, primarily caused by the lack of corresponding parametric knowledge in LLMs. An efficient solution is to enhance the knowledge of LLMs at test-time. Parametric RAG (PRAG) addresses this by embedding document into LLMs parameters to perform test-time knowledge enhancement, effectively reducing inference costs through offline training. However, its high training and storage costs, along with limited generalization ability, significantly restrict its practical adoption. To address these challenges, we propose Dynamic Parametric RAG (DyPRAG), a novel framework that leverages a lightweight parameter translator model to efficiently convert documents into parametric knowledge. DyPRAG not only reduces inference, training, and storage costs but also dynamically generates parametric knowledge, seamlessly enhancing the knowledge of LLMs and resolving knowledge conflicts in a plug-and-play manner at test-time. Extensive experiments on multiple datasets demonstrate the effectiveness and generalization capabilities of DyPRAG, offering a powerful and practical RAG paradigm which enables superior knowledge fusion and mitigates RAG hallucination in real-world applications. Our code is available at this https URL.

25 A Survey on Personalized Alignment -- The Missing Piece for Large Language Models in Real-World Applications

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, Wei Wu
https://arxiv.org/abs/arXiv:2503.17003
Large Language Models (LLMs) have demonstrated remarkable capabilities, yet their

transition to real-world applications reveals a critical limitation: the inability to adapt to individual preferences while maintaining alignment with universal human values. Current alignment techniques adopt a one-size-fits-all approach that fails to accommodate users' diverse backgrounds and needs. This paper presents the first comprehensive survey of personalized alignment-a paradigm that enables LLMs to adapt their behavior within ethical boundaries based on individual preferences. We propose a unified framework comprising preference memory management, personalized generation, and feedback-based alignment, systematically analyzing implementation approaches and evaluating their effectiveness across various scenarios. By examining current techniques, potential risks, and future challenges, this survey provides a structured foundation for developing more adaptable and ethically-aligned LLMs.

26 A Survey on Inference Engines for Large Language Models: Perspectives on Optimization and Efficiency

Sihyeong Park, Sungryeol Jeon, Chaelyn Lee, Seokhun Jeon, Byung-Soo Kim, Jemin Lee

https://arxiv.org/abs/arXiv:2505.01658

Large language models (LLMs) are widely applied in chatbots, code generators, and search engines. Workloads such as chain-of-thought, complex reasoning, and agent services significantly increase the inference cost by invoking the model repeatedly. Optimization methods such as parallelism, compression, and caching have been adopted to reduce costs, but the diverse service requirements make it hard to select the right method. Recently, specialized LLM inference engines have emerged as a key component for integrating the optimization methods into service-oriented infrastructures. However, a systematic study on inference engines is still lacking. This paper provides a comprehensive evaluation of 25 open-source and commercial inference engines. We examine each inference engine in terms of ease-of-use, ease-of-deployment, general-purpose support, scalability, and suitability for throughput- and latency-aware computation. Furthermore, we explore the design goals of each inference engine by investigating the optimization techniques it supports. In addition, we assess the ecosystem maturity of open source inference engines and handle the performance and cost policy of commercial solutions. We outline future research directions that include support for complex LLM-based services, support of various hardware, and enhanced security, offering practical guidance to researchers and developers in selecting and designing optimized LLM inference engines. We also provide a public repository to continually track developments in this fast-evolving field: this https URL

27 Colombian Waitresses y Jueces canadienses: Gender and Country Biases in Occupation Recommendations from LLMs

Elisa Forcada Rodríguez, Olatz Perez-de-Viñaspre, Jon Ander Campos, Dietrich Klakow, Vagrant Gautam

https://arxiv.org/abs/arXiv:2505.02456
One of the goals of fairness research in NLP is to measure and mitigate stereotypical biases that are propagated by NLP systems. However, such work tends to focus on single axes of bias (most often gender) and the English language. Addressing these limitations, we contribute the first study of multilingual intersecting country and gender biases, with a focus on occupation recommendations generated by large language models. We construct a benchmark of prompts in English, Spanish and German, where we systematically vary country and gender, using 25 countries and four pronoun sets. Then, we evaluate a suite of 5 Llama-based models on this benchmark, finding that LLMs encode significant gender and country biases. Notably, we find that even when models show parity for gender or country individually, intersectional occupational biases based on both country and gender persist. We also show that the prompting language significantly affects bias, and instruction-tuned models consistently demonstrate the lowest and most stable levels of bias. Our findings highlight the need for fairness researchers to use intersectional and multilingual lenses in their work.

28 ParaICL: Towards Parallel In-Context Learning

Xingxuan Li, Xuan-Phi Nguyen, Shafiq Joty, Lidong Bing
https://arxiv.org/abs/arXiv:2404.00570
Large language models (LLMs) have become the norm in natural language processing (NLP), excelling in few-shot in-context learning (ICL) with their remarkable abilities. Nonetheless, the success of ICL largely hinges on the choice of few-shot demonstration examples, making the selection process increasingly crucial. Existing methods have delved into optimizing the quantity and semantic similarity of these examples to improve ICL performances. However, our preliminary experiments indicate that the effectiveness of ICL is limited by the length of the input context. Moreover, varying combinations of few-shot demonstration examples can significantly boost accuracy across different test samples. To address this, we propose a novel method named parallel in-context learning (ParaICL) that effectively utilizes all demonstration examples without exceeding the manageable input context length. ParaICL employs parallel batching to distribute demonstration examples into different batches according to the semantic similarities of the questions in the demonstrations to the test question. It then computes normalized batch semantic scores for each batch. A weighted average semantic objective, constrained by adaptive plausibility, is applied to select the most appropriate tokens. Through extensive experiments, we validate the effectiveness of ParaICL and conduct ablation studies to underscore its design rationale. We further demonstrate that ParaICL can seamlessly integrate with existing methods.

29 R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, Liang Wang

https://arxiv.org/abs/arXiv:2505.02835

Multimodal Reward Models (MRMs) play a crucial role in enhancing the performance of Multimodal Large Language Models (MLLMs). While recent advancements have primarily focused on improving the model structure and training data of MRMs, there has been limited exploration into the effectiveness of long-term reasoning capabilities for reward modeling and how to activate these capabilities in MRMs. In this paper, we explore how Reinforcement Learning (RL) can be used to improve reward modeling. Specifically, we reformulate the reward modeling problem as a rule-based RL task. However, we observe that directly applying existing RL algorithms, such as Reinforce++, to reward modeling often leads to training instability or even collapse due to the inherent limitations of these algorithms. To address this issue, we propose the StableReinforce algorithm, which refines the training loss, advantage estimation strategy, and reward design of existing RL methods. These refinements result in more stable training dynamics and superior performance. To facilitate MRM training, we collect 200K preference data from diverse datasets. Our reward model, R1-Reward, trained using the StableReinforce algorithm on this dataset, significantly improves performance on multimodal reward modeling benchmarks. Compared to previous SOTA models, R1-Reward achieves a \$8.4\%\$ improvement on the VL Reward-Bench and a \$14.3\%\$ improvement on the Multimodal Reward Bench. Moreover, with more inference compute, R1-Reward's performance is further enhanced, highlighting the potential of RL algorithms in optimizing MRMs.

30 A Logical Fallacy-Informed Framework for Argument Generation

Luca Mouchel, Debjit Paul, Shaobo Cui, Robert West, Antoine Bosselut, Boi Faltings

https://arxiv.org/abs/arXiv:2408.03618

Despite the remarkable performance of Large Language Models (LLMs) in natural language processing tasks, they still struggle with generating logically sound arguments, resulting in potential risks such as spreading misinformation. To address this issue, we introduce FIPO, a fallacy-informed framework that leverages preference optimization methods to steer LLMs toward logically sound arguments. FIPO includes a classification loss, to capture the fine-grained information on fallacy types. Our results on argumentation datasets show that our method reduces the fallacy errors by up to 17.5%. Furthermore, our human evaluation results indicate that the quality of the generated arguments by our method significantly outperforms the fine-tuned baselines, as well as other preference optimization methods, such as DPO. These findings highlight the importance of ensuring models are aware of logical fallacies for effective argument generation. Our code is available at this http URL.

31 Analyzing Cognitive Differences Among Large Language Models through the Lens of Social Worldview

Jiatao Li, Yanheng Li, Xiaojun Wan
https://arxiv.org/abs/arXiv:2505.01967
Large Language Models (LLMs) have become integral to daily life, widely adopted in communication, decision-making, and information retrieval, raising critical questions about how these systems implicitly form and express socio-cognitive attitudes or "worldviews". While existing research extensively addresses demographic and ethical biases, broader dimensions-such as attitudes toward authority, equality, autonomy, and fate-remain under-explored. In this paper, we introduce the Social Worldview Taxonomy (SWT), a structured framework grounded in Cultural Theory, operationalizing four canonical worldviews (Hierarchy, Egalitarianism, Individualism, Fatalism) into measurable sub-dimensions. Using SWT, we empirically identify distinct and interpretable cognitive profiles across 28 diverse LLMs. Further, inspired by Social Referencing Theory, we experimentally demonstrate that explicit social cues systematically shape these cognitive attitudes, revealing both general response patterns and nuanced model-specific variations. Our findings enhance the interpretability of LLMs by revealing implicit socio-cognitive biases and their responsiveness to social feedback, thus guiding the development of more transparent and socially responsible language technologies.

32 Sailing AI by the Stars: A Survey of Learning from Rewards in Post-Training and Test-Time Scaling of Large Language Models

Xiaobao Wu
https://arxiv.org/abs/arXiv:2505.02686
Recent developments in Large Language Models (LLMs) have shifted from pre-training scaling to post-training and test-time scaling. Across these developments, a key unified paradigm has arisen: Learning from Rewards, where reward signals act as the guiding stars to steer LLM behavior. It has underpinned a wide range of prevalent techniques, such as reinforcement learning (in RLHF, DPO, and GRPO), reward-guided decoding, and post-hoc correction. Crucially, this paradigm enables the transition from passive learning from static data to active learning from dynamic feedback. This endows LLMs with aligned preferences and deep reasoning capabilities. In this survey, we present a comprehensive overview of the paradigm of learning from rewards. We categorize and analyze the strategies under this paradigm across training, inference, and post-inference stages. We further discuss the benchmarks for reward models and the primary applications. Finally we highlight the challenges and future directions. We maintain a paper collection at this https URL.

33 DECIDER: A Dual-System Rule-Controllable Decoding Framework for Language Generation

Chen Xu, Tian Lan, Yu Ji, Changlong Yu, Wei Wang, Jun Gao, Qunxi Dong, Kun Qian, Piji Li, Wei Bi, Bin Hu
https://arxiv.org/abs/arXiv:2403.01954
Constrained decoding approaches aim to control the meaning or style of text generated by the pre-trained large language models (LLMs or also PLMs) for various tasks at inference time. However, these methods often guide plausible continuations by greedily and explicitly selecting targets. Though fulfilling the task requirements, these methods may overlook certain general and natural logics that humans would implicitly follow towards such targets. Inspired by cognitive dual-process theory, in this work, we propose a novel decoding framework DECIDER where the base LLMs are equipped with a First-Order Logic (FOL) reasoner to express and evaluate the rules, along with a decision function that merges the outputs of both systems to guide the generation. Unlike previous constrained decodings, DECIDER transforms the encouragement of target-specific words into all words that satisfy several high-level rules, enabling us to programmatically integrate our logic into LLMs. Experiments on CommonGen and PersonaChat demonstrate that DECIDER effectively follows given FOL rules to guide LLMs in a more human-like and logic-controlled manner.

34 LUSIFER: Language Universal Space Integration for Enhanced Multilingual Embeddings with Large Language Models

Hieu Man, Nghia Trung Ngo, Viet Dac Lai, Ryan A. Rossi, Franck Dernoncourt, Thien Huu Nguyen
https://arxiv.org/abs/arXiv:2501.00874
Recent advancements in large language models (LLMs) based embedding models have established new state-of-the-art benchmarks for text embedding tasks, particularly in dense vector-based retrieval. However, these models predominantly focus on English, leaving multilingual embedding capabilities largely unexplored. To address this limitation, we present LUSIFER, a novel zero-shot approach that adapts LLM-based embedding models for multilingual tasks without requiring multilingual supervision. LUSIFER's architecture combines a multilingual encoder, serving as a language-universal learner, with an LLM-based embedding model optimized for embedding-specific tasks. These components are seamlessly integrated through a minimal set of trainable parameters that act as a connector, effectively transferring the multilingual encoder's language understanding capabilities to the specialized embedding model. Additionally, to comprehensively evaluate multilingual embedding performance, we introduce a new benchmark encompassing 5 primary embedding tasks, 123 diverse datasets, and coverage across 14 languages. Extensive experimental results demonstrate that LUSIFER significantly enhances the multilingual performance across various embedding tasks, particularly for medium and low-resource languages, without requiring explicit multilingual training data.

35 FairTranslate: An English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity

Fanny Jourdan, Yannick Chevalier, Cécile Favre
https://arxiv.org/abs/arXiv:2504.15941
Large Language Models (LLMs) are increasingly leveraged for translation tasks but often fall short when translating inclusive language -- such as texts containing the singular 'they' pronoun or otherwise reflecting fair linguistic protocols. Because these challenges span both computational and societal domains, it is imperative to critically evaluate how well LLMs handle inclusive translation with a well-founded framework.
This paper presents FairTranslate, a novel, fully human-annotated dataset designed to evaluate non-binary gender biases in machine translation systems from English to French. FairTranslate includes 2418 English-French sentence pairs related to occupations, annotated with rich metadata such as the stereotypical alignment of the occupation, grammatical gender indicator ambiguity, and the ground-truth gender label (male, female, or inclusive).
We evaluate four leading LLMs (Gemma2-2B, Mistral-7B, Llama3.1-8B, Llama3.3-70B) on this dataset under different prompting procedures. Our results reveal substantial biases in gender representation across LLMs, highlighting persistent challenges in achieving equitable outcomes in machine translation. These findings underscore the need for focused strategies and interventions aimed at ensuring fair and inclusive language usage in LLM-based translation systems.
We make the FairTranslate dataset publicly available on Hugging Face, and disclose the code for all experiments on GitHub.

36 Better Estimation of the KL Divergence Between Language Models

Afra Amini, Tim Vieira, Ryan Cotterell
https://arxiv.org/abs/arXiv:2504.10637
Estimating the Kullback--Leibler (KL) divergence between language models has many applications, e.g., reinforcement learning from human feedback (RLHF), interpretability, and knowledge distillation. However, computing the exact KL divergence between two arbitrary language models is intractable. Thus, practitioners often resort to the use of sampling-based estimators. While it is easy to fashion a simple Monte Carlo (MC) estimator that provides an unbiased estimate of the KL divergence between language models, this estimator notoriously suffers from high variance, and can even result in a negative estimate of the KL divergence, a non-negative quantity. In this paper, we introduce a Rao--Blackwellized estimator that is also unbiased and provably has variance less than or equal to that of the standard Monte Carlo estimator. In an empirical study on sentiment-controlled fine-tuning, we show that our estimator provides more stable

KL estimates and reduces variance substantially in practice. Additionally, we derive an analogous Rao--Blackwellized estimator of the gradient of the KL divergence, which leads to more stable training and produces models that more frequently appear on the Pareto frontier of reward vs. KL compared to the ones trained with the MC estimator of the gradient.

37 EMORL: Ensemble Multi-Objective Reinforcement Learning for Efficient and Flexible LLM Fine-Tuning

Lingxiao Kong (1), Cong Yang (2), Susanne Neufang (3), Oya Deniz Beyan (1,3), Zeyd Boukhers (1,3) ((1) Fraunhofer Institute for Applied Information Technology FIT, (2) Soochow University, (3) University Hospital of Cologne)

https://arxiv.org/abs/arXiv:2505.02579

Recent advances in reinforcement learning (RL) for large language model (LLM) fine-tuning show promise in addressing multi-objective tasks but still face significant challenges, including complex objective balancing, low training efficiency, poor scalability, and limited explainability. Leveraging ensemble learning principles, we introduce an Ensemble Multi-Objective RL (EMORL) framework that fine-tunes multiple models with individual objectives while optimizing their aggregation after the training to improve efficiency and flexibility. Our method is the first to aggregate the last hidden states of individual models, incorporating contextual information from multiple objectives. This approach is supported by a hierarchical grid search algorithm that identifies optimal weighted combinations. We evaluate EMORL on counselor reflection generation tasks, using text-scoring LLMs to evaluate the generations and provide rewards during RL fine-tuning. Through comprehensive experiments on the PAIR and Psych8k datasets, we demonstrate the advantages of EMORL against existing baselines: significantly lower and more stable training consumption ($17,529\pm 1,650$ data points and $6,573\pm 147.43$ seconds), improved scalability and explainability, and comparable performance across multiple objectives.

38 Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, Yongfeng Zhang

https://arxiv.org/abs/arXiv:2502.01563

Large language models (LLMs) have achieved remarkable success in contextual knowledge understanding. In this paper, we show that these concentrated massive values consistently emerge in specific regions of attention queries (Q) and keys (K) while not having such patterns in values (V) in various modern transformer-based LLMs (Q, K, and V mean the representations output by the query, key, and value layers respectively). Through extensive experiments, we further demonstrate that these massive values play

a critical role in interpreting contextual knowledge (knowledge obtained from the current context window) rather than in retrieving parametric knowledge stored within the model's parameters. Our further investigation of quantization strategies reveals that ignoring these massive values leads to a pronounced drop in performance on tasks requiring rich contextual understanding, aligning with our analysis. Finally, we trace the emergence of concentrated massive values and find that such concentration is caused by Rotary Positional Encoding (RoPE), which has appeared since the first layers. These findings shed new light on how Q and K operate in LLMs and offer practical insights for model design and optimization. The Code is Available at this https URL.

39 Mapping Trustworthiness in Large Language Models: A Bibliometric Analysis Bridging Theory to Practice

José Siqueira de Cerqueira, Kai-Kristian Kemell, Rebekah Rousi, Nannan Xi, Juho Hamari, Pekka Abrahamsson
https://arxiv.org/abs/arXiv:2503.04785
The rapid proliferation of Large Language Models (LLMs) has raised significant trustworthiness and ethical concerns. Despite the widespread adoption of LLMs across domains, there is still no clear consensus on how to define and operationalise trustworthiness. This study aims to bridge the gap between theoretical discussion and practical implementation by analysing research trends, definitions of trustworthiness, and practical techniques. We conducted a bibliometric mapping analysis of 2,006 publications from Web of Science (2019-2025) using the Bibliometrix, and manually reviewed 68 papers. We found a shift from traditional AI ethics discussion to LLM trustworthiness frameworks. We identified 18 different definitions of trust/trustworthiness, with transparency, explainability and reliability emerging as the most common dimensions. We identified 20 strategies to enhance LLM trustworthiness, with fine-tuning and retrieval-augmented generation (RAG) being the most prominent. Most of the strategies are developer-driven and applied during the post-training phase. Several authors propose fragmented terminologies rather than unified frameworks, leading to the risks of "ethics washing," where ethical discourse is adopted without a genuine regulatory commitment. Our findings highlight: persistent gaps between theoretical taxonomies and practical implementation, the crucial role of the developer in operationalising trust, and call for standardised frameworks and stronger regulatory measures to enable trustworthy and ethical deployment of LLMs.

40 The Illusion of Role Separation: Hidden Shortcuts in LLM Role Learning (and How to Fix Them)

Zihao Wang, Yibo Jiang, Jiahao Yu, Heqing Huang
https://arxiv.org/abs/arXiv:2505.00626
Large language models (LLMs) that integrate multiple input roles (e.g., system instructions, user queries, external tool outputs) are increasingly prevalent in practice.

Ensuring that the model accurately distinguishes messages from each role -- a concept we call \emph{role separation} -- is crucial for consistent multi-role behavior. Although recent work often targets state-of-the-art prompt injection defenses, it remains unclear whether such methods truly teach LLMs to differentiate roles or merely memorize known triggers. In this paper, we examine \emph{role-separation learning}: the process of teaching LLMs to robustly distinguish system and user tokens. Through a \emph{simple, controlled experimental framework}, we find that fine-tuned models often rely on two proxies for role identification: (1) task type exploitation, and (2) proximity to begin-of-text. Although data augmentation can partially mitigate these shortcuts, it generally leads to iterative patching rather than a deeper fix. To address this, we propose reinforcing \emph{invariant signals} that mark role boundaries by adjusting token-wise cues in the model's input encoding. In particular, manipulating position IDs helps the model learn clearer distinctions and reduces reliance on superficial proxies. By focusing on this mechanism-centered perspective, our work illuminates how LLMs can more reliably maintain consistent multi-role behavior without merely memorizing known prompts or triggers.

41 A Survey on Progress in LLM Alignment from the Perspective of Reward Design

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, Usman Naseem

https://arxiv.org/abs/arXiv:2505.02666

The alignment of large language models (LLMs) with human values and intentions represents a core challenge in current AI research, where reward mechanism design has become a critical factor in shaping model behavior. This study conducts a comprehensive investigation of reward mechanisms in LLM alignment through a systematic theoretical framework, categorizing their development into three key phases: (1) feedback (diagnosis), (2) reward design (prescription), and (3) optimization (treatment). Through a four-dimensional analysis encompassing construction basis, format, expression, and granularity, this research establishes a systematic classification framework that reveals evolutionary trends in reward modeling. The field of LLM alignment faces several persistent challenges, while recent advances in reward design are driving significant paradigm shifts. Notable developments include the transition from reinforcement learning-based frameworks to novel optimization paradigms, as well as enhanced capabilities to address complex alignment scenarios involving multimodal integration and concurrent task coordination. Finally, this survey outlines promising future research directions for LLM alignment through innovative reward design strategies.

42 Bielik 11B v2 Technical Report

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, Remigiusz Kinas

https://arxiv.org/abs/arXiv:2505.02410
We present Bielik 11B v2, a state-of-the-art language model optimized for Polish text processing. Built on the Mistral 7B v0.2 architecture and scaled to 11B parameters using depth up-scaling, this model demonstrates exceptional performance across Polish language benchmarks while maintaining strong cross-lingual capabilities. We introduce two key technical innovations: Weighted Instruction Cross-Entropy Loss, which optimizes learning across diverse instruction types by assigning quality-based weights to training examples, and Adaptive Learning Rate, which dynamically adjusts based on context length. Comprehensive evaluation across multiple benchmarks demonstrates that Bielik 11B v2 outperforms many larger models, including those with 2-6 times more parameters, and significantly surpasses other specialized Polish language models on tasks ranging from linguistic understanding to complex reasoning. The model's parameter efficiency and extensive quantization options enable deployment across various hardware configurations, advancing Polish language AI capabilities and establishing new benchmarks for resource-efficient language modeling in less-represented languages.

43 The Effectiveness of Large Language Models in Transforming Unstructured Text to Standardized Formats

William Brach, Kristián Košťál, Michal Ries
https://arxiv.org/abs/arXiv:2503.02650
The exponential growth of unstructured text data presents a fundamental challenge in modern data management and information retrieval. While Large Language Models (LLMs) have shown remarkable capabilities in natural language processing, their potential to transform unstructured text into standardized, structured formats remains largely unexplored - a capability that could revolutionize data processing workflows across industries. This study breaks new ground by systematically evaluating LLMs' ability to convert unstructured recipe text into the structured Cooklang format. Through comprehensive testing of four models (GPT-4o, GPT-4o-mini, Llama3.1:70b, and Llama3.1:8b), an innovative evaluation approach is introduced that combines traditional metrics (WER, ROUGE-L, TER) with specialized metrics for semantic element identification. Our experiments reveal that GPT-4o with few-shot prompting achieves breakthrough performance (ROUGE-L: 0.9722, WER: 0.0730), demonstrating for the first time that LLMs can reliably transform domain-specific unstructured text into structured formats without extensive training. Although model performance generally scales with size, we uncover surprising potential in smaller models like Llama3.1:8b for optimization through targeted fine-tuning. These findings open new possibilities for automated structured data generation across various domains, from medical records to technical documentation, potentially transforming the way organizations process and utilize unstructured information.

44 Noise Augmented Fine Tuning for Mitigating Hallucinations in Large Language Models

Afshin Khadangi, Amir Sartipi, Igor Tchappi, Ramin Bahmani
https://arxiv.org/abs/arXiv:2504.03302
Large language models (LLMs) often produce inaccurate or misleading content-hallucinations. To address this challenge, we introduce Noise-Augmented Fine-Tuning (NoiseFiT), a novel framework that leverages adaptive noise injection based on the signal-to-noise ratio (SNR) to enhance model robustness. In particular, NoiseFiT selectively perturbs layers identified as either high-SNR (more robust) or low-SNR (potentially underregularized) using a dynamically scaled Gaussian noise. We further propose a hybrid loss that combines standard cross-entropy, soft cross-entropy, and consistency regularization to ensure stable and accurate outputs under noisy training conditions. Our theoretical analysis shows that adaptive noise injection is both unbiased and variance-preserving, providing strong guarantees for convergence in expectation. Empirical results on multiple test and benchmark datasets demonstrate that NoiseFiT significantly reduces hallucination rates, often improving or matching baseline performance in key tasks. These findings highlight the promise of noise-driven strategies for achieving robust, trustworthy language modeling without incurring prohibitive computational overhead. Given the comprehensive and detailed nature of our experiments, we have publicly released the fine-tuning logs, benchmark evaluation artifacts, and source code online at W&B, Hugging Face, and GitHub, respectively, to foster further research, accessibility and reproducibility.

45 Position: Enough of Scaling LLMs! Lets Focus on Downscaling

Ayan Sengupta, Yash Goel, Tanmoy Chakraborty
https://arxiv.org/abs/arXiv:2505.00985
We challenge the dominant focus on neural scaling laws and advocate for a paradigm shift toward downscaling in the development of large language models (LLMs). While scaling laws have provided critical insights into performance improvements through increasing model and dataset size, we emphasize the significant limitations of this approach, particularly in terms of computational inefficiency, environmental impact, and deployment constraints. To address these challenges, we propose a holistic framework for downscaling LLMs that seeks to maintain performance while drastically reducing resource demands. This paper outlines practical strategies for transitioning away from traditional scaling paradigms, advocating for a more sustainable, efficient, and accessible approach to LLM development.

46 SymPlanner: Deliberate Planning in Language Models with Symbolic Representation

Siheng Xiong, Jieyu Zhou, Zhangding Liu, Yusen Su
https://arxiv.org/abs/arXiv:2505.01479
Planning remains a core challenge for language models (LMs), particularly in domains that require coherent multi-step action sequences grounded in external constraints. We introduce SymPlanner, a novel framework that equips LMs with structured planning capabilities by interfacing them with a symbolic environment that serves as an explicit world model. Rather than relying purely on natural language reasoning, SymPlanner grounds the planning process in a symbolic state space, where a policy model proposes actions and a symbolic environment deterministically executes and verifies their effects. To enhance exploration and improve robustness, we introduce Iterative Correction (IC), which refines previously proposed actions by leveraging feedback from the symbolic environment to eliminate invalid decisions and guide the model toward valid alternatives. Additionally, Contrastive Ranking (CR) enables fine-grained comparison of candidate plans by evaluating them jointly. We evaluate SymPlanner on PlanBench, demonstrating that it produces more coherent, diverse, and verifiable plans than pure natural language baselines.

47 A New HOPE: Domain-agnostic Automatic Evaluation of Text Chunking

Henrik Brådland, Morten Goodwin, Per-Arne Andersen, Alexander S. Nossum, Aditya Gupta
https://arxiv.org/abs/arXiv:2505.02171
Document chunking fundamentally impacts Retrieval-Augmented Generation (RAG) by determining how source materials are segmented before indexing. Despite evidence that Large Language Models (LLMs) are sensitive to the layout and structure of retrieved data, there is currently no framework to analyze the impact of different chunking methods. In this paper, we introduce a novel methodology that defines essential characteristics of the chunking process at three levels: intrinsic passage properties, extrinsic passage properties, and passages-document coherence. We propose HOPE (Holistic Passage Evaluation), a domain-agnostic, automatic evaluation metric that quantifies and aggregates these characteristics. Our empirical evaluations across seven domains demonstrate that the HOPE metric correlates significantly ($p > 0.13$) with various RAG performance indicators, revealing contrasts between the importance of extrinsic and intrinsic properties of passages. Semantic independence between passages proves essential for system performance with a performance gain of up to 56.2% in factual correctness and 21.1% in answer correctness. On the contrary, traditional assumptions about maintaining concept unity within passages show minimal impact. These findings provide actionable insights for optimizing chunking strategies, thus improving RAG system design to produce more factually correct responses.

48 Bielik v3 Small: Technical Report

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, Adrian Gwoździej

https://arxiv.org/abs/arXiv:2505.02550

We introduce Bielik v3, a series of parameter-efficient generative text models (1.5B and 4.5B) optimized for Polish language processing. These models demonstrate that smaller, well-optimized architectures can achieve performance comparable to much larger counterparts while requiring substantially fewer computational resources. Our approach incorporates several key innovations: a custom Polish tokenizer (APT4) that significantly improves token efficiency, Weighted Instruction Cross-Entropy Loss to balance learning across instruction types, and Adaptive Learning Rate that dynamically adjusts based on training progress. Trained on a meticulously curated corpus of 292 billion tokens spanning 303 million documents, these models excel across multiple benchmarks, including the Open PL LLM Leaderboard, Complex Polish Text Understanding Benchmark, Polish EQ-Bench, and Polish Medical Leaderboard. The 4.5B parameter model achieves results competitive with models 2-3 times its size, while the 1.5B model delivers strong performance despite its extremely compact profile. These advances establish new benchmarks for parameter-efficient language modeling in less-represented languages, making high-quality Polish language AI more accessible for resource-constrained applications.

49 Identifying Legal Holdings with LLMs: A Systematic Study of Performance, Scale, and Memorization

Chuck Arvin

https://arxiv.org/abs/arXiv:2505.02172

As large language models (LLMs) continue to advance in capabilities, it is essential to assess how they perform on established benchmarks. In this study, we present a suite of experiments to assess the performance of modern LLMs (ranging from 3B to 90B+ parameters) on CaseHOLD, a legal benchmark dataset for identifying case holdings. Our experiments demonstrate "scaling effects" - performance on this task improves with model size, with more capable models like GPT4o and AmazonNovaPro achieving macro F1 scores of 0.744 and 0.720 respectively. These scores are competitive with the best published results on this dataset, and do not require any technically sophisticated model training, fine-tuning or few-shot prompting. To ensure that these strong results are not due to memorization of judicial opinions contained in the training data, we develop and utilize a novel citation anonymization test that preserves semantic meaning while ensuring case names and citations are fictitious. Models maintain strong performance under these conditions (macro F1 of 0.728), suggesting the performance is not due to rote memorization. These findings demonstrate both the promise and current limitations of LLMs for legal tasks with important implications for the development and measurement of automated legal analytics and legal benchmarks.

50 Towards Safer Pretraining: Analyzing and Filtering Harmful Content in Webscale datasets for Responsible LLMs

Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, Parag Agrawal

https://arxiv.org/abs/arXiv:2505.02009

Large language models (LLMs) have become integral to various real-world applications, leveraging massive, web-sourced datasets like Common Crawl, C4, and FineWeb for pretraining. While these datasets provide linguistic data essential for high-quality natural language generation, they often contain harmful content, such as hate speech, misinformation, and biased narratives. Training LLMs on such unfiltered data risks perpetuating toxic behaviors, spreading misinformation, and amplifying societal biases which can undermine trust in LLM-driven applications and raise ethical concerns about their use. This paper presents a large-scale analysis of inappropriate content across these datasets, offering a comprehensive taxonomy that categorizes harmful webpages into Topical and Toxic based on their intent. We also introduce a prompt evaluation dataset, a high-accuracy Topical and Toxic Prompt (TTP), and a transformer-based model (HarmFormer) for content filtering. Additionally, we create a new multi-harm open-ended toxicity benchmark (HAVOC) and provide crucial insights into how models respond to adversarial toxic inputs. Upon publishing, we will also opensource our model signal on the entire C4 dataset. Our work offers insights into ensuring safer LLM pretraining and serves as a resource for Responsible AI (RAI) compliance.

51 Structured Prompting and Feedback-Guided Reasoning with LLMs for Data Interpretation

Amit Rath

https://arxiv.org/abs/arXiv:2505.01636

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and task generalization. However, their application to structured data analysis remains fragile due to inconsistencies in schema interpretation, misalignment between user intent and model output, and limited mechanisms for self-correction when failures occur. This paper introduces the STROT Framework (Structured Task Reasoning and Output Transformation), a method for structured prompting and feedback-driven transformation logic generation aimed at improving the reliability and semantic alignment of LLM-based analytical workflows. STROT begins with lightweight schema introspection and sample-based field classification, enabling dynamic context construction that captures both the structure and statistical profile of the input data. This contextual information is embedded in structured prompts that guide the model toward generating task-specific, interpretable outputs. To address common failure modes in complex queries, STROT incorporates a refinement mechanism in which the model iteratively revises its outputs based on execution feedback and validation signals. Unlike conventional approaches that rely on static prompts or single-shot inference, STROT treats the LLM as a reasoning agent embedded within a controlled analysis loop -- capable of adjusting its output trajectory through planning and correction. The result is a robust and reproducible framework for reasoning over

structured data with LLMs, applicable to diverse data exploration and analysis tasks where interpretability, stability, and correctness are essential.

52 On the effectiveness of Large Language Models in the mechanical design domain

Daniele Grandi, Fabian Riquelme
https://arxiv.org/abs/arXiv:2505.01559
In this work, we seek to understand the performance of large language models in the mechanical engineering domain. We leverage the semantic data found in the ABC dataset, specifically the assembly names that designers assigned to the overall assemblies, and the individual semantic part names that were assigned to each part. After pre-processing the data we developed two unsupervised tasks to evaluate how different model architectures perform on domain-specific data: a binary sentence-pair classification task and a zero-shot classification task. We achieved a 0.62 accuracy for the binary sentence-pair classification task with a fine-tuned model that focuses on fighting over-fitting: 1) modifying learning rates, 2) dropout values, 3) Sequence Length, and 4) adding a multi-head attention layer. Our model on the zero-shot classification task outperforms the baselines by a wide margin, and achieves a top-1 classification accuracy of 0.386. The results shed some light on the specific failure modes that arise when learning from language in this domain.

53 Helping Large Language Models Protect Themselves: An Enhanced Filtering and Summarization System

Sheikh Samit Muhaimin, Spyridon Mastorakis
https://arxiv.org/abs/arXiv:2505.01315
The recent growth in the use of Large Language Models has made them vulnerable to sophisticated adversarial assaults, manipulative prompts, and encoded malicious inputs. Existing countermeasures frequently necessitate retraining models, which is computationally costly and impracticable for deployment. Without the need for retraining or fine-tuning, this study presents a unique defense paradigm that allows LLMs to recognize, filter, and defend against adversarial or malicious inputs on their own. There are two main parts to the suggested framework: (1) A prompt filtering module that uses sophisticated Natural Language Processing (NLP) techniques, including zero-shot classification, keyword analysis, and encoded content detection (e.g. base64, hexadecimal, URL encoding), to detect, decode, and classify harmful inputs; and (2) A summarization module that processes and summarizes adversarial research literature to give the LLM context-aware defense knowledge. This approach strengthens LLMs' resistance to adversarial exploitation by fusing text extraction, summarization, and harmful prompt analysis. According to experimental results, this integrated technique has a 98.71% success rate in identifying harmful patterns, manipulative language structures, and encoded prompts. By employing a modest amount of adversarial research literature as

context, the methodology also allows the model to react correctly to harmful inputs with a larger percentage of jailbreak resistance and refusal rate. While maintaining the quality of LLM responses, the framework dramatically increases LLM's resistance to hostile misuse, demonstrating its efficacy as a quick and easy substitute for time-consuming, retraining-based defenses.

54 Knowing You Don't Know: Learning When to Continue Search in Multi-round RAG through Self-Practicing

Diji Yang, Linda Zeng, Jinmeng Rao, Yi Zhang
https://arxiv.org/abs/arXiv:2505.02811
Retrieval Augmented Generation (RAG) has shown strong capability in enhancing language models' knowledge and reducing AI generative hallucinations, driving its widespread use. However, complex tasks requiring multi-round retrieval remain challenging, and early attempts tend to be overly optimistic without a good sense of self-skepticism. Current multi-round RAG systems may continue searching even when enough information has already been retrieved, or they may provide incorrect answers without having sufficient information or knowledge. Existing solutions either require large amounts of expensive human-labeled process supervision data or lead to subpar performance.
This paper aims to address these limitations by introducing a new framework, \textbf{SIM-RAG}, to explicitly enhance RAG systems' self-awareness and multi-round retrieval capabilities. To train SIM-RAG, we first let a RAG system self-practice multi-round retrieval, augmenting existing question-answer pairs with intermediate inner monologue reasoning steps to generate synthetic training data. For each pair, the system may explore multiple retrieval paths, which are labeled as successful if they reach the correct answer and unsuccessful otherwise. Using this data, we train a lightweight information sufficiency Critic. At inference time, the Critic evaluates whether the RAG system has retrieved sufficient information at each round, guiding retrieval decisions and improving system-level self-awareness through in-context reinforcement learning.
Experiments across multiple prominent RAG benchmarks show that SIM-RAG is an effective multi-round RAG solution. Furthermore, this framework is system-efficient, adding a lightweight component to RAG without requiring modifications to existing LLMs or search engines, and data-efficient, eliminating the need for costly human-annotated mid-step retrieval process supervision data.

55 CHORUS: Zero-shot Hierarchical Retrieval and Orchestration for Generating Linear Programming Code

Tasnim Ahmed, Salimur Choudhury
https://arxiv.org/abs/arXiv:2505.01485
Linear Programming (LP) problems aim to find the optimal solution to an objective under constraints. These problems typically require domain knowledge, mathematical

skills, and programming ability, presenting significant challenges for non-experts. This study explores the efficiency of Large Language Models (LLMs) in generating solver-specific LP code. We propose CHORUS, a retrieval-augmented generation (RAG) framework for synthesizing Gurobi-based LP code from natural language problem statements. CHORUS incorporates a hierarchical tree-like chunking strategy for theoretical contents and generates additional metadata based on code examples from documentation to facilitate self-contained, semantically coherent retrieval. Two-stage retrieval approach of CHORUS followed by cross-encoder reranking further ensures contextual relevance. Finally, expertly crafted prompt and structured parser with reasoning steps improve code generation performance significantly. Experiments on the NL4Opt-Code benchmark show that CHORUS improves the performance of open-source LLMs such as Llama3.1 (8B), Llama3.3 (70B), Phi4 (14B), Deepseek-r1 (32B), and Qwen2.5-coder (32B) by a significant margin compared to baseline and conventional RAG. It also allows these open-source LLMs to outperform or match the performance of much stronger baselines-GPT3.5 and GPT4 while requiring far fewer computational resources. Ablation studies further demonstrate the importance of expert prompting, hierarchical chunking, and structured reasoning.

56 Attention Mechanisms Perspective: Exploring LLM Processing of Graph-Structured Data

Zhong Guan, Likang Wu, Hongke Zhao, Ming He, Jianpin Fan
https://arxiv.org/abs/arXiv:2505.02130
Attention mechanisms are critical to the success of large language models (LLMs), driving significant advancements in multiple fields. However, for graph-structured data, which requires emphasis on topological connections, they fall short compared to message-passing mechanisms on fixed links, such as those employed by Graph Neural Networks (GNNs). This raises a question: "Does attention fail for graphs in natural language settings?" Motivated by these observations, we embarked on an empirical study from the perspective of attention mechanisms to explore how LLMs process graph-structured data. The goal is to gain deeper insights into the attention behavior of LLMs over graph structures. We uncovered unique phenomena regarding how LLMs apply attention to graph-structured data and analyzed these findings to improve the modeling of such data by LLMs. The primary findings of our research are: 1) While LLMs can recognize graph data and capture text-node interactions, they struggle to model inter-node relationships within graph structures due to inherent architectural constraints. 2) The attention distribution of LLMs across graph nodes does not align with ideal structural patterns, indicating a failure to adapt to graph topology nuances. 3) Neither fully connected attention nor fixed connectivity is optimal; each has specific limitations in its application scenarios. Instead, intermediate-state attention windows improve LLM training performance and seamlessly transition to fully connected windows during inference. Source code: \href{this https URL}{LLM4Exploration}

57 Improving Phishing Email Detection Performance of Small Large Language Models

Zijie Lin, Zikang Liu, Hanbo Fan
https://arxiv.org/abs/arXiv:2505.00034
Large language models(LLMs) have demonstrated remarkable performance on many natural language processing(NLP) tasks and have been employed in phishing email detection research. However, in current studies, well-performing LLMs typically contain billions or even tens of billions of parameters, requiring enormous computational resources. To reduce computational costs, we investigated the effectiveness of small-parameter LLMs for phishing email detection. These LLMs have around 3 billion parameters and can run on consumer-grade GPUs. However, small LLMs often perform poorly in phishing email detection task. To address these issues, we designed a set of methods including Prompt Engineering, Explanation Augmented Fine-tuning, and Model Ensemble to improve phishing email detection capabilities of small LLMs. We validated the effectiveness of our approach through experiments, significantly improving both accuracy and F1 score on the SpamAssassin and CEAS_08 datasets. Furthermore, the fine-tuned models demonstrated strong transferability, achieving robust performance across multiple unseen phishing datasets, outperforming traditional baselines and approaching standard-sized LLMs.

58 Bye-bye, Bluebook? Automating Legal Procedure with Large Language Models

Matthew Dahl
https://arxiv.org/abs/arXiv:2505.02763
Legal practice requires careful adherence to procedural rules. In the United States, few are more complex than those found in The Bluebook: A Uniform System of Citation. Compliance with this system's 500+ pages of byzantine formatting instructions is the raison d'etre of thousands of student law review editors and the bete noire of lawyers everywhere. To evaluate whether large language models (LLMs) are able to adhere to the procedures of such a complicated system, we construct an original dataset of 866 Bluebook tasks and test flagship LLMs from OpenAI, Anthropic, Google, Meta, and DeepSeek. We show (1) that these models produce fully compliant Bluebook citations only 69%-74% of the time and (2) that in-context learning on the Bluebook's underlying system of rules raises accuracy only to 77%. These results caution against using off-the-shelf LLMs to automate aspects of the law where fidelity to procedure is paramount.

59 Data Augmentation With Back translation for Low Resource languages: A case of English and Luganda

Richard Kimera, Dongnyeong Heo, Daniela N. Rim, Heeyoul Choi
https://arxiv.org/abs/arXiv:2505.02463
In this paper,we explore the application of Back translation (BT) as a semi-supervised technique to enhance Neural Machine Translation(NMT) models for the English-Luganda language pair, specifically addressing the challenges faced by low-resource languages. The purpose of our study is to demonstrate how BT can mitigate the scarcity of bilingual data by generating synthetic data from monolingual corpora. Our methodology involves developing custom NMT models using both publicly available and web-crawled data, and applying Iterative and Incremental Back translation techniques. We strategically select datasets for incremental back translation across multiple small datasets, which is a novel element of our approach. The results of our study show significant improvements, with translation performance for the English-Luganda pair exceeding previous benchmarks by more than 10 BLEU score units across all translation directions. Additionally, our evaluation incorporates comprehensive assessment metrics such as SacreBLEU, ChrF2, and TER, providing a nuanced understanding of translation quality. The conclusion drawn from our research confirms the efficacy of BT when strategically curated datasets are utilized, establishing new performance benchmarks and demonstrating the potential of BT in enhancing NMT models for low-resource languages.

60 Parameter-Efficient Transformer Embeddings

Henry Ndubuaku, Mouad Talhi
https://arxiv.org/abs/arXiv:2505.02266
Embedding layers in transformer-based NLP models typically account for the largest share of model parameters, scaling with vocabulary size but not yielding performance gains proportional to scale. We propose an alternative approach in which token embedding vectors are first generated deterministically, directly from the token IDs using a Fourier expansion of their normalized values, followed by a lightweight multilayer perceptron (MLP) that captures higher-order interactions. We train standard transformers and our architecture on natural language inference tasks (SNLI and MNLI), and evaluate zero-shot performance on sentence textual similarity (STS-B). Our results demonstrate that the proposed method achieves competitive performance using significantly fewer parameters, trains faster, and operates effectively without the need for dropout. This proof-of-concept study highlights the potential for scalable, memory-efficient language models and motivates further large-scale experimentation based on our findings.

61 LLaMA-Omni2: LLM-based Real-time Spoken Chatbot with Autoregressive Streaming Speech Synthesis

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, Yang Feng
https://arxiv.org/abs/arXiv:2505.02625
Real-time, intelligent, and natural speech interaction is an essential part of the next-generation human-computer interaction. Recent advancements have showcased the potential of building intelligent spoken chatbots based on large language models (LLMs). In this paper, we introduce LLaMA-Omni 2, a series of speech language models (SpeechLMs) ranging from 0.5B to 14B parameters, capable of achieving high-quality real-time speech interaction. LLaMA-Omni 2 is built upon the Qwen2.5 series models, integrating a speech encoder and an autoregressive streaming speech decoder. Despite being trained on only 200K multi-turn speech dialogue samples, LLaMA-Omni 2 demonstrates strong performance on several spoken question answering and speech instruction following benchmarks, surpassing previous state-of-the-art SpeechLMs like GLM-4-Voice, which was trained on millions of hours of speech data.

62 Unlearning Sensitive Information in Multimodal LLMs: Benchmark and Attack-Defense Evaluation

Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, Mohit Bansal
https://arxiv.org/abs/arXiv:2505.01456
LLMs trained on massive datasets may inadvertently acquire sensitive information such as personal details and potentially harmful content. This risk is further heightened in multimodal LLMs as they integrate information from multiple modalities (image and text). Adversaries can exploit this knowledge through multimodal prompts to extract sensitive details. Evaluating how effectively MLLMs can forget such information (targeted unlearning) necessitates the creation of high-quality, well-annotated image-text pairs. While prior work on unlearning has focused on text, multimodal unlearning remains underexplored. To address this gap, we first introduce a multimodal unlearning benchmark, UnLOK-VQA (Unlearning Outside Knowledge VQA), as well as an attack-and-defense framework to evaluate methods for deleting specific multimodal knowledge from MLLMs. We extend a visual question-answering dataset using an automated pipeline that generates varying-proximity samples for testing generalization and specificity, followed by manual filtering for maintaining high quality. We then evaluate six defense objectives against seven attacks (four whitebox, three blackbox), including a novel whitebox method leveraging interpretability of hidden states. Our results show multimodal attacks outperform text- or image-only ones, and that the most effective defense removes answer information from internal model states. Additionally, larger models exhibit greater post-editing robustness, suggesting that scale enhances safety. UnLOK-VQA provides a rigorous benchmark for advancing unlearning in MLLMs.

63 ResearchCodeAgent: An LLM Multi-Agent System for Automated Codification of Research Methodologies

Shubham Gandhi, Dhruv Shah, Manasi Patwardhan, Lovekesh Vig, Gautam Shroff

https://arxiv.org/abs/arXiv:2504.20117

In this paper we introduce ResearchCodeAgent, a novel multi-agent system leveraging large language models (LLMs) agents to automate the codification of research methodologies described in machine learning literature. The system bridges the gap between high-level research concepts and their practical implementation, allowing researchers auto-generating code of existing research papers for benchmarking or building on top-of existing methods specified in the literature with availability of partial or complete starter code. ResearchCodeAgent employs a flexible agent architecture with a comprehensive action suite, enabling context-aware interactions with the research environment. The system incorporates a dynamic planning mechanism, utilizing both short and long-term memory to adapt its approach iteratively. We evaluate ResearchCodeAgent on three distinct machine learning tasks with distinct task complexity and representing different parts of the ML pipeline: data augmentation, optimization, and data batching. Our results demonstrate the system's effectiveness and generalizability, with 46.9% of generated code being high-quality and error-free, and 25% showing performance improvements over baseline implementations. Empirical analysis shows an average reduction of 57.9% in coding time compared to manual implementation. We observe higher gains for more complex tasks. ResearchCodeAgent represents a significant step towards automating the research implementation process, potentially accelerating the pace of machine learning research.

64 Optimizing LLMs for Resource-Constrained Environments: A Survey of Model Compression Techniques

Sanjay Surendranath Girija, Shashank Kapoor, Lakshit Arora, Dipen Pradhan, Aman Raj, Ankit Shetgaonkar

https://arxiv.org/abs/arXiv:2505.02309

Large Language Models (LLMs) have revolutionized many areas of artificial intelligence (AI), but their substantial resource requirements limit their deployment on mobile and edge devices. This survey paper provides a comprehensive overview of techniques for compressing LLMs to enable efficient inference in resource-constrained environments. We examine three primary approaches: Knowledge Distillation, Model Quantization, and Model Pruning. For each technique, we discuss the underlying principles, present different variants, and provide examples of successful applications. We also briefly discuss complementary techniques such as mixture-of-experts and early-exit strategies. Finally, we highlight promising future directions, aiming to provide a valuable resource for both researchers and practitioners seeking to optimize LLMs for edge deployment.

65 MoxE: Mixture of xLSTM Experts with Entropy-Aware Routing for Efficient Language Modeling

Abdoul Majid O. Thiombiano, Brahim Hnich, Ali Ben Mrad, Mohamed Wiem Mkaouer

https://arxiv.org/abs/arXiv:2505.01459

This paper introduces MoxE, a novel architecture that synergistically combines the Extended Long Short-Term Memory (xLSTM) with the Mixture of Experts (MoE) framework to address critical scalability and efficiency challenges in large language models (LLMs). The proposed method effectively leverages xLSTM's innovative memory structures while strategically introducing sparsity through MoE to substantially reduce computational overhead. At the heart of our approach is a novel entropy-based routing mechanism, designed to dynamically route tokens to specialized experts, thereby ensuring efficient and balanced resource utilization. This entropy awareness enables the architecture to effectively manage both rare and common tokens, with mLSTM blocks being favored to handle rare tokens. To further enhance generalization, we introduce a suite of auxiliary losses, including entropy-based and group-wise balancing losses, ensuring robust performance and efficient training. Theoretical analysis and empirical evaluations rigorously demonstrate that MoxE achieves significant efficiency gains and enhanced effectiveness compared to existing approaches, marking a notable advancement in scalable LLM architectures.

66 From Human Judgements to Predictive Models: Unravelling Acceptability in Code-Mixed Sentences

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Ponnurangam Kumaraguru, Manish Shrivastava

https://arxiv.org/abs/arXiv:2405.05572

Current computational approaches for analysing or generating code-mixed sentences do not explicitly model "naturalness" or "acceptability" of code-mixed sentences, but rely on training corpora to reflect distribution of acceptable code-mixed sentences. Modelling human judgement for the acceptability of code-mixed text can help in distinguishing natural code-mixed text and enable quality-controlled generation of code-mixed text. To this end, we construct Cline - a dataset containing human acceptability judgements for English-Hindi~(en-hi) code-mixed text. Cline is the largest of its kind with 16,642 sentences, consisting of samples sourced from two sources: synthetically generated code-mixed text and samples collected from online social media. Our analysis establishes that popular code-mixing metrics such as CMI, Number of Switch Points, Burstines, which are used to filter/curate/compare code-mixed corpora have low correlation with human acceptability judgements, underlining the necessity of our dataset. Experiments using Cline demonstrate that simple Multilayer Perceptron (MLP) models when trained solely using code-mixing metrics as features are outperformed by fine-tuned pre-trained Multilingual Large Language Models (MLLMs). Specifically, among Encoder models XLM-Roberta and Bernice outperform IndicBERT across different configurations. Among Encoder-Decoder models, mBART performs better than

mT5, however Encoder-Decoder models are not able to outperform Encoder-only models. Decoder-only models perform the best when compared to all other MLLMS, with Llama 3.2 - 3B models outperforming similarly sized Qwen, Phi models. Comparison with zero and fewshot capabilitites of ChatGPT show that MLLMs fine-tuned on larger data outperform ChatGPT, providing scope for improvement in code-mixed tasks. Zero-shot transfer from En-Hi to En-Te acceptability judgments are better than random baselines.

67 ELECTRA and GPT-4o: Cost-Effective Partners for Sentiment Analysis

James P. Beno
https://arxiv.org/abs/arXiv:2501.00062
Bidirectional transformers excel at sentiment analysis, and Large Language Models (LLM) are effective zero-shot learners. Might they perform better as a team? This paper explores collaborative approaches between ELECTRA and GPT-4o for three-way sentiment classification. We fine-tuned (FT) four models (ELECTRA Base/Large, GPT-4o/4o-mini) using a mix of reviews from Stanford Sentiment Treebank (SST) and DynaSent. We provided input from ELECTRA to GPT as: predicted label, probabilities, and retrieved examples. Sharing ELECTRA Base FT predictions with GPT-4o-mini significantly improved performance over either model alone (82.50 macro F1 vs. 79.14 ELECTRA Base FT, 79.41 GPT-4o-mini) and yielded the lowest cost/performance ratio (\$0.12/F1 point). However, when GPT models were fine-tuned, including predictions decreased performance. GPT-4o FT-M was the top performer (86.99), with GPT-4o-mini FT close behind (86.70) at much less cost (\$0.38 vs. \$1.59/F1 point). Our results show that augmenting prompts with predictions from fine-tuned encoders is an efficient way to boost performance, and a fine-tuned GPT-4o-mini is nearly as good as GPT-4o FT at 76% less cost. Both are affordable options for projects with limited resources.

68 Prompt-Based Cost-Effective Evaluation and Operation of ChatGPT as a Computer Programming Teaching Assistant

Marc Ballestero-Ribó, Daniel Ortiz-Martínez
https://arxiv.org/abs/arXiv:2501.17176
The dream of achieving a student-teacher ratio of 1:1 is closer than ever thanks to the emergence of large language models (LLMs). One potential application of these models in the educational field would be to provide feedback to students in university introductory programming courses, so that a student struggling to solve a basic implementation problem could seek help from an LLM available 24/7. This article focuses on studying three aspects related to such an application. First, the performance of two well-known models, GPT-3.5T and GPT-4T, in providing feedback to students is evaluated. The empirical results showed that GPT-4T performs much better than GPT-3.5T, however,

it is not yet ready for use in a real-world scenario. This is due to the possibility of generating incorrect information that potential users may not always be able to detect. Second, the article proposes a carefully designed prompt using in-context learning techniques that allows automating important parts of the evaluation process, as well as providing a lower bound for the fraction of feedbacks containing incorrect information, saving time and effort. This was possible because the resulting feedback has a programmatically analyzable structure that incorporates diagnostic information about the LLM's performance in solving the requested task. Third, the article also suggests a possible strategy for implementing a practical learning tool based on LLMs, which is rooted on the proposed prompting techniques. This strategy opens up a whole range of interesting possibilities from a pedagogical perspective.

69 AutoLibra: Agent Metric Induction from Open-Ended Feedback

Hao Zhu, Phil Cuvin, Xinkai Yu, Charlotte Ka Yee Yan, Jason Zhang, Diyi Yang

https://arxiv.org/abs/arXiv:2505.02820

Agents are predominantly evaluated and optimized via task success metrics, which are coarse, rely on manual design from experts, and fail to reward intermediate emergent behaviors. We propose AutoLibra, a framework for agent evaluation, that transforms open-ended human feedback, e.g., "If you find that the button is disabled, don't click it again", or "This agent has too much autonomy to decide what to do on its own", into metrics for evaluating fine-grained behaviors in agent trajectories. AutoLibra accomplishes this by grounding feedback to an agent's behavior, clustering similar positive and negative behaviors, and creating concrete metrics with clear definitions and concrete examples, which can be used for prompting LLM-as-a-Judge as evaluators. We further propose two meta-metrics to evaluate the alignment of a set of (induced) metrics with open feedback: "coverage" and "redundancy". Through optimizing these meta-metrics, we experimentally demonstrate AutoLibra's ability to induce more concrete agent evaluation metrics than the ones proposed in previous agent evaluation benchmarks and discover new metrics to analyze agents. We also present two applications of AutoLibra in agent improvement: First, we show that AutoLibra-induced metrics serve as better prompt-engineering targets than the task success rate on a wide range of text game tasks, improving agent performance over baseline by a mean of 20%. Second, we show that AutoLibra can iteratively select high-quality fine-tuning data for web navigation agents. Our results suggest that AutoLibra is a powerful task-agnostic tool for evaluating and improving language agents.

70 Applying LLMs to Active Learning: Towards Cost-Efficient Cross-Task Text Classification without Manually Labeled Data

Yejian Zhang, Shingo Takada
https://arxiv.org/abs/arXiv:2502.16892
Machine learning-based classifiers have been used for text classification, such as sentiment analysis, news classification, and toxic comment classification. However, supervised machine learning models often require large amounts of labeled data for training, and manual annotation is both labor-intensive and requires domain-specific knowledge, leading to relatively high annotation costs. To address this issue, we propose an approach that integrates large language models (LLMs) into an active learning framework, achieving high cross-task text classification performance without the need for any manually labeled data. Furthermore, compared to directly applying GPT for classification tasks, our approach retains over 93% of its classification performance while requiring only approximately 6% of the computational time and monetary cost, effectively balancing performance and resource efficiency. These findings provide new insights into the efficient utilization of LLMs and active learning algorithms in text classification tasks, paving the way for their broader application.

71 Think on your Feet: Adaptive Thinking via Reinforcement Learning for Social Agents

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, Wenji Mao
https://arxiv.org/abs/arXiv:2505.02156
Effective social intelligence simulation requires language agents to dynamically adjust reasoning depth, a capability notably absent in current approaches. While existing methods either lack this kind of reasoning capability or enforce uniform long chain-of-thought reasoning across all scenarios, resulting in excessive token usage and inappropriate social simulation. In this paper, we propose $\textbf{A}$daptive $\textbf{M}$ode $\textbf{L}$earning ($\textbf{AML}$) that strategically selects from four thinking modes (intuitive reaction $\rightarrow$ deep contemplation) based on real-time context. Our framework's core innovation, the $\textbf{A}$daptive $\textbf{M}$ode $\textbf{P}$olicy $\textbf{O}$ptimization ($\textbf{AMPO}$) algorithm, introduces three key advancements over existing methods: (1) Multi-granular thinking mode design, (2) Context-aware mode switching across social interaction, and (3) Token-efficient reasoning via depth-adaptive processing. Extensive experiments on social intelligence tasks confirm that AML achieves 15.6% higher task performance than state-of-the-art methods. Notably, our method outperforms GRPO by 7.0% with 32.8% shorter reasoning chains. These results demonstrate that context-sensitive thinking mode selection, as implemented in AMPO, enables more human-like adaptive reasoning than GRPO's fixed-depth approach

72 PIPA: A Unified Evaluation Protocol for Diagnosing Interactive Planning Agents

Takyoung Kim, Janvijay Singh, Shuhaib Mehri, Emre Can Acikgoz, Sagnik Mukherjee, Nimet Beyza Bozdag, Sumuk Shashidhar, Gokhan Tur, Dilek Hakkani-Tür

https://arxiv.org/abs/arXiv:2505.01592

The growing capabilities of large language models (LLMs) in instruction-following and context-understanding lead to the era of agents with numerous applications. Among these, task planning agents have become especially prominent in realistic scenarios involving complex internal pipelines, such as context understanding, tool management, and response generation. However, existing benchmarks predominantly evaluate agent performance based on task completion as a proxy for overall effectiveness. We hypothesize that merely improving task completion is misaligned with maximizing user satisfaction, as users interact with the entire agentic process and not only the end result. To address this gap, we propose PIPA, a unified evaluation protocol that conceptualizes the behavioral process of interactive task planning agents within a partially observable Markov Decision Process (POMDP) paradigm. The proposed protocol offers a comprehensive assessment of agent performance through a set of atomic evaluation criteria, allowing researchers and practitioners to diagnose specific strengths and weaknesses within the agent's decision-making pipeline. Our analyses show that agents excel in different behavioral stages, with user satisfaction shaped by both outcomes and intermediate behaviors. We also highlight future directions, including systems that leverage multiple agents and the limitations of user simulators in task planning.

73 Almost AI, Almost Human: The Challenge of Detecting AI-Polished Writing

Shoumik Saha, Soheil Feizi

https://arxiv.org/abs/arXiv:2502.15666

The growing use of large language models (LLMs) for text generation has led to widespread concerns about AI-generated content detection. However, an overlooked challenge is AI-polished text, where human-written content undergoes subtle refinements using AI tools. This raises a critical question: should minimally polished text be classified as AI-generated? Such classification can lead to false plagiarism accusations and misleading claims about AI prevalence in online content. In this study, we systematically evaluate twelve state-of-the-art AI-text detectors using our AI-Polished-Text Evaluation (APT-Eval) dataset, which contains 14.7K samples refined at varying AI-involvement levels. Our findings reveal that detectors frequently flag even minimally polished text as AI-generated, struggle to differentiate between degrees of AI involvement, and exhibit biases against older and smaller models. These limitations highlight the urgent need for more nuanced detection methodologies.

74 Automatic Proficiency Assessment in L2 English Learners

Armita Mohammadi, Alessandro Lameiras Koerich, Laureano Moro-Velazquez, Patrick Cardinal

https://arxiv.org/abs/arXiv:2505.02615

Second language proficiency (L2) in English is usually perceptually evaluated by English teachers or expert evaluators, with the inherent intra- and inter-rater variability. This paper explores deep learning techniques for comprehensive L2 proficiency assessment, addressing both the speech signal and its correspondent transcription. We analyze spoken proficiency classification prediction using diverse architectures, including 2D CNN, frequency-based CNN, ResNet, and a pretrained wav2vec 2.0 model. Additionally, we examine text-based proficiency assessment by fine-tuning a BERT language model within resource constraints. Finally, we tackle the complex task of spontaneous dialogue assessment, managing long-form audio and speaker interactions through separate applications of wav2vec 2.0 and BERT models. Results from experiments on EFCamDat and ANGLISH datasets and a private dataset highlight the potential of deep learning, especially the pretrained wav2vec 2.0 model, for robust automated L2 proficiency evaluation.

75 ReplaceMe: Network Simplification via Layer Pruning and Linear Transformations

Dmitriy Shopkhoev, Ammar Ali, Magauiya Zhussip, Valentin Malykh, Stamatios Lefkimmiatis, Nikos Komodakis, Sergey Zagoruyko

https://arxiv.org/abs/arXiv:2505.02819

We introduce ReplaceMe, a generalized training-free depth pruning method that effectively replaces transformer blocks with a linear operation, while maintaining high performance for low compression ratios. In contrast to conventional pruning approaches that require additional training or fine-tuning, our approach requires only a small calibration dataset that is used to estimate a linear transformation to approximate the pruned blocks. This estimated linear mapping can be seamlessly merged with the remaining transformer blocks, eliminating the need for any additional network parameters. Our experiments show that ReplaceMe consistently outperforms other training-free approaches and remains highly competitive with state-of-the-art pruning methods that involve extensive retraining/fine-tuning and architectural modifications. Applied to several large language models (LLMs), ReplaceMe achieves up to 25% pruning while retaining approximately 90% of the original model's performance on open benchmarks - without any training or healing steps, resulting in minimal computational overhead (see Fig.1). We provide an open-source library implementing ReplaceMe alongside several state-of-the-art depth pruning techniques, available at this repository.

76 SEval-Ex: A Statement-Level Framework for Explainable Summarization Evaluation

Tanguy Herserant, Vincent Guigue
https://arxiv.org/abs/arXiv:2505.02235
Evaluating text summarization quality remains a critical challenge in Natural Language Processing. Current approaches face a trade-off between performance and interpretability. We present SEval-Ex, a framework that bridges this gap by decomposing summarization evaluation into atomic statements, enabling both high performance and explainability. SEval-Ex employs a two-stage pipeline: first extracting atomic statements from text source and summary using LLM, then a matching between generated statements. Unlike existing approaches that provide only summary-level scores, our method generates detailed evidence for its decisions through statement-level alignments. Experiments on the SummEval benchmark demonstrate that SEval-Ex achieves state-of-the-art performance with 0.580 correlation on consistency with human consistency judgments, surpassing GPT-4 based evaluators (0.521) while maintaining interpretability. Finally, our framework shows robustness against hallucination.

77 AD-LLM: Benchmarking Large Language Models for Anomaly Detection

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, Xia Hu, Yue Zhao
https://arxiv.org/abs/arXiv:2412.11142
Anomaly detection (AD) is an important machine learning task with many real-world uses, including fraud detection, medical diagnosis, and industrial monitoring. Within natural language processing (NLP), AD helps detect issues like spam, misinformation, and unusual user activity. Although large language models (LLMs) have had a strong impact on tasks such as text generation and summarization, their potential in AD has not been studied enough. This paper introduces AD-LLM, the first benchmark that evaluates how LLMs can help with NLP anomaly detection. We examine three key tasks: (i) zero-shot detection, using LLMs' pre-trained knowledge to perform AD without tasks-specific training; (ii) data augmentation, generating synthetic data and category descriptions to improve AD models; and (iii) model selection, using LLMs to suggest unsupervised AD models. Through experiments with different datasets, we find that LLMs can work well in zero-shot AD, that carefully designed augmentation methods are useful, and that explaining model selection for specific datasets remains challenging. Based on these results, we outline six future research directions on LLMs for AD.

78 Enhancing the Learning Experience: Using Vision-Language Models to Generate Questions for Educational Videos

Markos Stamatakis, Joshua Berger, Christian Wartena, Ralph Ewerth, Anett Hoppe

https://arxiv.org/abs/arXiv:2505.01790

Web-based educational videos offer flexible learning opportunities and are becoming increasingly popular. However, improving user engagement and knowledge retention remains a challenge. Automatically generated questions can activate learners and support their knowledge acquisition. Further, they can help teachers and learners assess their understanding. While large language and vision-language models have been employed in various tasks, their application to question generation for educational videos remains underexplored. In this paper, we investigate the capabilities of current vision-language models for generating learning-oriented questions for educational video content. We assess (1) out-of-the-box models' performance; (2) fine-tuning effects on content-specific question generation; (3) the impact of different video modalities on question quality; and (4) in a qualitative study, question relevance, answerability, and difficulty levels of generated questions. Our findings delineate the capabilities of current vision-language models, highlighting the need for fine-tuning and addressing challenges in question diversity and relevance. We identify requirements for future multimodal datasets and outline promising research directions.

79 JTCSE: Joint Tensor-Modulus Constraints and Cross-Attention for Unsupervised Contrastive Learning of Sentence Embeddings

Tianyu Zong, Hongzhu Yi, Bingkang Shi, Yuanxiang Wang, Jungang Xu

https://arxiv.org/abs/arXiv:2505.02366

Unsupervised contrastive learning has become a hot research topic in natural language processing. Existing works usually aim at constraining the orientation distribution of the representations of positive and negative samples in the high-dimensional semantic space in contrastive learning, but the semantic representation tensor possesses both modulus and orientation features, and the existing works ignore the modulus feature of the representations and cause insufficient contrastive learning. % Therefore, we firstly propose a training objective that aims at modulus constraints on the semantic representation tensor, to strengthen the alignment between the positive samples in contrastive learning. Therefore, we first propose a training objective that is designed to impose modulus constraints on the semantic representation tensor, to strengthen the alignment between positive samples in contrastive learning. Then, the BERT-like model suffers from the phenomenon of sinking attention, leading to a lack of attention to CLS tokens that aggregate semantic information. In response, we propose a cross-attention structure among the twin-tower ensemble models to enhance the model's attention to CLS token and optimize the quality of CLS Pooling. Combining the above two motivations, we propose a new \textbf{J}oint \textbf{T}ensor representation modulus constraint and \textbf{C}ross-attention unsupervised contrastive learn-

ing \textbf{S}entence \textbf{E}mbedding representation framework JTCSE, which we evaluate in seven semantic text similarity computation tasks, and the experimental results show that JTCSE's twin-tower ensemble model and single-tower distillation model outperform the other baselines and become the current SOTA. In addition, we have conducted an extensive zero-shot downstream task evaluation, which shows that JTCSE outperforms other baselines overall on more than 130 tasks.

80 High-Fidelity Pseudo-label Generation by Large Language Models for Training Robust Radiology Report Classifiers

Brian Wong, Kaito Tanaka
https://arxiv.org/abs/arXiv:2505.01693
Automated labeling of chest X-ray reports is essential for enabling downstream tasks such as training image-based diagnostic models, population health studies, and clinical decision support. However, the high variability, complexity, and prevalence of negation and uncertainty in these free-text reports pose significant challenges for traditional Natural Language Processing methods. While large language models (LLMs) demonstrate strong text understanding, their direct application for large-scale, efficient labeling is limited by computational cost and speed. This paper introduces DeBERTa-RAD, a novel two-stage framework that combines the power of state-of-the-art LLM pseudo-labeling with efficient DeBERTa-based knowledge distillation for accurate and fast chest X-ray report labeling. We leverage an advanced LLM to generate high-quality pseudo-labels, including certainty statuses, for a large corpus of reports. Subsequently, a DeBERTa-Base model is trained on this pseudo-labeled data using a tailored knowledge distillation strategy. Evaluated on the expert-annotated MIMIC-500 benchmark, DeBERTa-RAD achieves a state-of-the-art Macro F1 score of 0.9120, significantly outperforming established rule-based systems, fine-tuned transformer models, and direct LLM inference, while maintaining a practical inference speed suitable for high-throughput applications. Our analysis shows particular strength in handling uncertain findings. This work demonstrates a promising path to overcome data annotation bottlenecks and achieve high-performance medical text processing through the strategic combination of LLM capabilities and efficient student models trained via distillation.

81 A Comprehensive Analysis for Visual Object Hallucination in Large Vision-Language Models

Liqiang Jing, Guiming Hardy Chen, Ehsan Aghazadeh, Xin Eric Wang, Xinya Du
https://arxiv.org/abs/arXiv:2505.01958
Large Vision-Language Models (LVLMs) demonstrate remarkable capabilities in multimodal tasks, but visual object hallucination remains a persistent issue. It refers to scenarios where models generate inaccurate visual object-related information based on the query input, potentially leading to misinformation and concerns about safety and

reliability. Previous works focus on the evaluation and mitigation of visual hallucinations, but the underlying causes have not been comprehensively investigated. In this paper, we analyze each component of LLaVA-like LVLMs -- the large language model, the vision backbone, and the projector -- to identify potential sources of error and their impact. Based on our observations, we propose methods to mitigate hallucination for each problematic component. Additionally, we developed two hallucination benchmarks: QA-VisualGenome, which emphasizes attribute and relation hallucinations, and QA-FB15k, which focuses on cognition-based hallucinations.

82 EgoNormia: Benchmarking Physical Social Norm Understanding

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, Diyi Yang
https://arxiv.org/abs/arXiv:2502.20490
Human activity is moderated by norms. However, machines are often trained without explicit supervision on norm understanding and reasoning, particularly when norms are physically- or socially-grounded. To improve and evaluate the normative reasoning capability of vision-language models (VLMs), we present \dataset{} $\|\epsilon\|$, consisting of 1,853 challenging, multi-stage MCQ questions based on ego-centric videos of human interactions, evaluating both the prediction and justification of normative actions. The normative actions encompass seven categories: safety, privacy, proxemics, politeness, cooperation, coordination/proactivity, and communication/legibility. To compile this dataset at scale, we propose a novel pipeline leveraging video sampling, automatic answer generation, filtering, and human validation. Our work demonstrates that current state-of-the-art vision-language models lack robust norm understanding, scoring a maximum of 54\% on \dataset{} (versus a human bench of 92\%). Our analysis of performance in each dimension highlights the significant risks of safety, privacy, and the lack of collaboration and communication capability when applied to real-world agents. We additionally show that through a retrieval-based generation (RAG) method, it is possible to use \dataset{} to enhance normative reasoning in VLMs.

83 LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric (Yifan)Cao, Yossi Matias, Avinatan Hassidim, Dale R Webster, Yun Liu, Sho Fujiwara, Peggy Bui, Quang Duong
https://arxiv.org/abs/arXiv:2505.01980
Information on the web, such as scientific publications and Wikipedia, often surpasses

users' reading level. To help address this, we used a self-refinement approach to develop a LLM capability for minimally lossy text simplification. To validate our approach, we conducted a randomized study involving 4563 participants and 31 texts spanning 6 broad subject areas: PubMed (biomedical scientific articles), biology, law, finance, literature/philosophy, and aerospace/computer science. Participants were randomized to viewing original or simplified texts in a subject area, and answered multiple-choice questions (MCQs) that tested their comprehension of the text. The participants were also asked to provide qualitative feedback such as task difficulty. Our results indicate that participants who read the simplified text answered more MCQs correctly than their counterparts who read the original text (3.9% absolute increase, $p<0.05$). This gain was most striking with PubMed (14.6%), while more moderate gains were observed for finance (5.5%), aerospace/computer science (3.8%) domains, and legal (3.5%). Notably, the results were robust to whether participants could refer back to the text while answering MCQs. The absolute accuracy decreased by up to ~9% for both original and simplified setups where participants could not refer back to the text, but the ~4% overall improvement persisted. Finally, participants' self-reported perceived ease based on a simplified NASA Task Load Index was greater for those who read the simplified text (absolute change on a 5-point scale 0.33, $p<0.05$). This randomized study, involving an order of magnitude more participants than prior works, demonstrates the potential of LLMs to make complex information easier to understand. Our work aims to enable a broader audience to better learn and make use of expert knowledge available on the web, improving information accessibility.

84 Generative Sign-description Prompts with Multi-positive Contrastive Learning for Sign Language Recognition

Siyu Liang, Yunan Li, Wentian Xin, Huizhou Chen, Xujie Liu, Kang Liu, Qiguang Miao

https://arxiv.org/abs/arXiv:2505.02304

Sign language recognition (SLR) faces fundamental challenges in creating accurate annotations due to the inherent complexity of simultaneous manual and non-manual signals. To the best of our knowledge, this is the first work to integrate generative large language models (LLMs) into SLR tasks. We propose a novel Generative Sign-description Prompts Multi-positive Contrastive learning (GSP-MC) method that leverages retrieval-augmented generation (RAG) with domain-specific LLMs, incorporating multi-step prompt engineering and expert-validated sign language corpora to produce precise multipart descriptions. The GSP-MC method also employs a dual-encoder architecture to bidirectionally align hierarchical skeleton features with multiple text descriptions (global, synonym, and part level) through probabilistic matching. Our approach combines global and part-level losses, optimizing KL divergence to ensure robust alignment across all relevant text-skeleton pairs while capturing both sign-level semantics and detailed part dynamics. Experiments demonstrate state-of-the-art performance against existing methods on the Chinese SLR500 (reaching 97.1%) and Turkish AUTSL datasets (97.07% accuracy). The method's cross-lingual effectiveness highlight its potential for developing inclusive communication technologies.

85 Voila: Voice-Language Foundation Models for Real-Time Autonomous Interaction and Voice Role-Play

Yemin Shi, Yu Shu, Siwei Dong, Guangyi Liu, Jaward Sesay, Jingwen Li, Zhiting Hu

https://arxiv.org/abs/arXiv:2505.02707

A voice AI agent that blends seamlessly into daily life would interact with humans in an autonomous, real-time, and emotionally expressive manner. Rather than merely reacting to commands, it would continuously listen, reason, and respond proactively, fostering fluid, dynamic, and emotionally resonant interactions. We introduce Voila, a family of large voice-language foundation models that make a step towards this vision. Voila moves beyond traditional pipeline systems by adopting a new end-to-end architecture that enables full-duplex, low-latency conversations while preserving rich vocal nuances such as tone, rhythm, and emotion. It achieves a response latency of just 195 milliseconds, surpassing the average human response time. Its hierarchical multi-scale Transformer integrates the reasoning capabilities of large language models (LLMs) with powerful acoustic modeling, enabling natural, persona-aware voice generation -- where users can simply write text instructions to define the speaker's identity, tone, and other characteristics. Moreover, Voila supports over one million pre-built voices and efficient customization of new ones from brief audio samples as short as 10 seconds. Beyond spoken dialogue, Voila is designed as a unified model for a wide range of voice-based applications, including automatic speech recognition (ASR), Text-to-Speech (TTS), and, with minimal adaptation, multilingual speech translation. Voila is fully open-sourced to support open research and accelerate progress toward next-generation human-machine interactions.

86 Ensemble Kalman filter for uncertainty in human language comprehension

Diksha Bhandari, Alessandro Lopopolo, Milena Rabovsky, Sebastian Reich

https://arxiv.org/abs/arXiv:2505.02590

Artificial neural networks (ANNs) are widely used in modeling sentence processing but often exhibit deterministic behavior, contrasting with human sentence comprehension, which manages uncertainty during ambiguous or unexpected inputs. This is exemplified by reversal anomalies-sentences with unexpected role reversals that challenge syntax and semantics-highlighting the limitations of traditional ANN models, such as the Sentence Gestalt (SG) Model. To address these limitations, we propose a Bayesian framework for sentence comprehension, applying an extension of the ensemble Kalman filter (EnKF) for Bayesian inference to quantify uncertainty. By framing language comprehension as a Bayesian inverse problem, this approach enhances the SG model's ability to reflect human sentence processing with respect to the representation of uncertainty. Numerical experiments and comparisons with maximum likelihood estimation (MLE) demonstrate that Bayesian methods improve uncertainty representation, enabling the

model to better approximate human cognitive processing when dealing with linguistic ambiguities.

87 Efficient Shapley Value-based Non-Uniform Pruning of Large Language Models

Chuan Sun, Han Yu, Lizhen Cui
https://arxiv.org/abs/arXiv:2505.01731
Pruning large language models (LLMs) is a promising solution for reducing model sizes and computational complexity while preserving performance. Traditional layer-wise pruning methods often adopt a uniform sparsity approach across all layers, which leads to suboptimal performance due to the varying significance of individual transformer layers within the model not being accounted for. To this end, we propose the \underline{S}hapley \underline{V}alue-based \underline{N}on-\underline{U}niform \underline{P}runing (\methodname{}) method for LLMs. This approach quantifies the contribution of each transformer layer to the overall model performance, enabling the assignment of tailored pruning budgets to different layers to retain critical parameters. To further improve efficiency, we design the Sliding Window-based Shapley Value approximation method. It substantially reduces computational overhead compared to exact SV calculation methods. Extensive experiments on various LLMs including LLaMA-v1, LLaMA-v2 and OPT demonstrate the effectiveness of the proposed approach. The results reveal that non-uniform pruning significantly enhances the performance of pruned models. Notably, \methodname{} achieves a reduction in perplexity (PPL) of 18.01\% and 19.55\% on LLaMA-7B and LLaMA-13B, respectively, compared to SparseGPT at 70\% sparsity.

88 An overview of artificial intelligence in computer-assisted language learning

Anisia Katinskaia
https://arxiv.org/abs/arXiv:2505.02032
Computer-assisted language learning -- CALL -- is an established research field. We review how artificial intelligence can be applied to support language learning and teaching. The need for intelligent agents that assist language learners and teachers is increasing: the human teacher's time is a scarce and costly resource, which does not scale with growing demand. Further factors contribute to the need for CALL: pandemics and increasing demand for distance learning, migration of large populations, the need for sustainable and affordable support for learning, etc. CALL systems are made up of many components that perform various functions, and AI is applied to many different aspects in CALL, corresponding to their own expansive research areas. Most of what we find in the research literature and in practical use are prototypes or partial implementations -- systems that perform some aspects of the overall desired functionality. Complete solutions -- most of them commercial -- are few, because they require massive

resources. Recent advances in AI should result in improvements in CALL, yet there is a lack of surveys that focus on AI in the context of this research field. This paper aims to present a perspective on the AI methods that can be employed for language learning from a position of a developer of a CALL system. We also aim to connect work from different disciplines, to build bridges for interdisciplinary work.

89 Impact of Noisy Supervision in Foundation Model Learning

Hao Chen, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, Bhiksha Raj, Jindong Wang

https://arxiv.org/abs/arXiv:2403.06869

Foundation models are usually pre-trained on large-scale datasets and then adapted to downstream tasks through tuning. However, the large-scale pre-training datasets, often inaccessible or too expensive to handle, can contain label noise that may adversely affect the generalization of the model and pose unexpected risks. This paper stands out as the first work to comprehensively understand and analyze the nature of noise in pre-training datasets and then effectively mitigate its impacts on downstream tasks. Specifically, through extensive experiments of fully-supervised and image-text contrastive pre-training on synthetic noisy ImageNet-1K, YFCC15M, and CC12M datasets, we demonstrate that, while slight noise in pre-training can benefit in-domain (ID) performance, where the training and testing data share a similar distribution, it always deteriorates out-of-domain (OOD) performance, where training and testing distributions are significantly different. These observations are agnostic to scales of pre-training datasets, pre-training noise types, model architectures, pre-training objectives, downstream tuning methods, and downstream applications. We empirically ascertain that the reason behind this is that the pre-training noise shapes the feature space differently. We then propose a tuning method (NMTune) to affine the feature space to mitigate the malignant effect of noise and improve generalization, which is applicable in both parameter-efficient and black-box tuning manners. We additionally conduct extensive experiments on popular vision and language models, including APIs, which are supervised and self-supervised pre-trained on realistic noisy data for evaluation. Our analysis and results demonstrate the importance of this novel and fundamental research direction, which we term as Noisy Model Learning.

90 Predicting Movie Hits Before They Happen with LLMs

Shaghayegh Agah, Yejin Kim, Neeraj Sharma, Mayur Nankani, Kevin Foley, H. Howie Huang, Sardar Hamidian

https://arxiv.org/abs/arXiv:2505.02693

Addressing the cold-start issue in content recommendation remains a critical ongoing challenge. In this work, we focus on tackling the cold-start problem for movies on a large entertainment platform. Our primary goal is to forecast the popularity of cold-start movies using Large Language Models (LLMs) leveraging movie metadata. This method could be integrated into retrieval systems within the personalization pipeline

or could be adopted as a tool for editorial teams to ensure fair promotion of potentially overlooked movies that may be missed by traditional or algorithmic solutions. Our study validates the effectiveness of this approach compared to established baselines and those we developed.

91 AdaParse: An Adaptive Parallel PDF Parsing and Resource Scaling Engine

Carlo Siebenschuh, Kyle Hippe, Ozan Gokdemir, Alexander Brace, Arham Khan, Khalid Hossain, Yadu Babuji, Nicholas Chia, Venkatram Vishwanath, Rick Stevens, Arvind Ramanathan, Ian Foster, Robert Underwood
https://arxiv.org/abs/arXiv:2505.01435
Language models for scientific tasks are trained on text from scientific publications, most distributed as PDFs that require parsing. PDF parsing approaches range from inexpensive heuristics (for simple documents) to computationally intensive ML-driven systems (for complex or degraded ones). The choice of the "best" parser for a particular document depends on its computational cost and the accuracy of its output. To address these issues, we introduce an Adaptive Parallel PDF Parsing and Resource Scaling Engine (AdaParse), a data-driven strategy for assigning an appropriate parser to each document. We enlist scientists to select preferred parser outputs and incorporate this information through direct preference optimization (DPO) into AdaParse, thereby aligning its selection process with human judgment. AdaParse then incorporates hardware requirements and predicted accuracy of each parser to orchestrate computational resources efficiently for large-scale parsing campaigns. We demonstrate that AdaParse, when compared to state-of-the-art parsers, improves throughput by $17\times$ while still achieving comparable accuracy (0.2 percent better) on a benchmark set of 1000 scientific documents. AdaParse's combination of high accuracy and parallel scalability makes it feasible to parse large-scale scientific document corpora to support the development of high-quality, trillion-token-scale text datasets. The implementation is available at this https URL

92 Using Knowledge Graphs to harvest datasets for efficient CLIP model training

Simon Ging, Sebastian Walter, Jelena Bratulić, Johannes Dienert, Hannah Bast, Thomas Brox
https://arxiv.org/abs/arXiv:2505.02746
Training high-quality CLIP models typically requires enormous datasets, which limits the development of domain-specific models -- especially in areas that even the largest CLIP models do not cover well -- and drives up training costs. This poses challenges for scientific research that needs fine-grained control over the training procedure of CLIP models. In this work, we show that by employing smart web search strategies enhanced with knowledge graphs, a robust CLIP model can be trained from scratch

with considerably less data. Specifically, we demonstrate that an expert foundation model for living organisms can be built using just 10M images. Moreover, we introduce EntityNet, a dataset comprising 33M images paired with 46M text descriptions, which enables the training of a generic CLIP model in significantly reduced time.

93 Exploring new Approaches for Information Retrieval through Natural Language Processing

Manak Raj, Nidhi Mishra
https://arxiv.org/abs/arXiv:2505.02199
This review paper explores recent advancements and emerging approaches in Information Retrieval (IR) applied to Natural Language Processing (NLP). We examine traditional IR models such as Boolean, vector space, probabilistic, and inference network models, and highlight modern techniques including deep learning, reinforcement learning, and pretrained transformer models like BERT. We discuss key tools and libraries - Lucene, Anserini, and Pyserini - for efficient text indexing and search. A comparative analysis of sparse, dense, and hybrid retrieval methods is presented, along with applications in web search engines, cross-language IR, argument mining, private information retrieval, and hate speech detection. Finally, we identify open challenges and future research directions to enhance retrieval accuracy, scalability, and ethical considerations.

94 SMUTF: Schema Matching Using Generative Tags and Hybrid Features

Yu Zhang, Mei Di, Haozheng Luo, Chenwei Xu, Richard Tzong-Han Tsai
https://arxiv.org/abs/arXiv:2402.01685
We introduce SMUTF (Schema Matching Using Generative Tags and Hybrid Features), a unique approach for large-scale tabular data schema matching (SM), which assumes that supervised learning does not affect performance in open-domain tasks, thereby enabling effective cross-domain matching. This system uniquely combines rule-based feature engineering, pre-trained language models, and generative large language models. In an innovative adaptation inspired by the Humanitarian Exchange Language, we deploy "generative tags" for each data column, enhancing the effectiveness of SM. SMUTF exhibits extensive versatility, working seamlessly with any pre-existing pre-trained embeddings, classification methods, and generative models.
Recognizing the lack of extensive, publicly available datasets for SM, we have created and open-sourced the HDXSM dataset from the public humanitarian data. We believe this to be the most exhaustive SM dataset currently available. In evaluations across various public datasets and the novel HDXSM dataset, SMUTF demonstrated exceptional performance, surpassing existing state-of-the-art models in terms of accuracy and efficiency, and improving the F1 score by 11.84% and the AUC of ROC by 5.08%. Code is available at this https URL.

95 LecEval: An Automated Metric for Multimodal Knowledge Acquisition in Multimedia Learning

Joy Lim Jia Yin, Daniel Zhang-Li, Jifan Yu, Haoxuan Li, Shangqing Tu, Yuanchun Wang, Zhiyuan Liu, Huiqin Liu, Lei Hou, Juanzi Li, Bin Xu
https://arxiv.org/abs/arXiv:2505.02078
Evaluating the quality of slide-based multimedia instruction is challenging. Existing methods like manual assessment, reference-based metrics, and large language model evaluators face limitations in scalability, context capture, or bias. In this paper, we introduce LecEval, an automated metric grounded in Mayer's Cognitive Theory of Multimedia Learning, to evaluate multimodal knowledge acquisition in slide-based learning. LecEval assesses effectiveness using four rubrics: Content Relevance (CR), Expressive Clarity (EC), Logical Structure (LS), and Audience Engagement (AE). We curate a large-scale dataset of over 2,000 slides from more than 50 online course videos, annotated with fine-grained human ratings across these rubrics. A model trained on this dataset demonstrates superior accuracy and adaptability compared to existing metrics, bridging the gap between automated and human assessments. We release our dataset and toolkits at this https URL.

96 Large Language Models as Carriers of Hidden Messages

Jakub Hoscilowicz, Pawel Popiolek, Jan Rudkowski, Jedrzej Bieniasz, Artur Janicki
https://arxiv.org/abs/arXiv:2406.02481
Simple fine-tuning can embed hidden text into large language models (LLMs), which is revealed only when triggered by a specific query. Applications include LLM fingerprinting, where a unique identifier is embedded to verify licensing compliance, and steganography, where the LLM carries hidden messages disclosed through a trigger query.

Our work demonstrates that embedding hidden text via fine-tuning, although seemingly secure due to the vast number of potential triggers, is vulnerable to extraction through analysis of the LLM's output decoding process. We introduce an extraction attack called Unconditional Token Forcing (UTF), which iteratively feeds tokens from the LLM's vocabulary to reveal sequences with high token probabilities, indicating hidden text candidates. We also present Unconditional Token Forcing Confusion (UTFC), a defense paradigm that makes hidden text resistant to all known extraction attacks without degrading the general performance of LLMs compared to standard fine-tuning. UTFC has both benign (improving LLM fingerprinting) and malign applications (using LLMs to create covert communication channels).

97 CAMOUFLAGE: Exploiting Misinformation Detection Systems Through LLM-driven Adversarial Claim Transformation

Mazal Bethany, Nishant Vishwamitra, Cho-Yu Jason Chiang, Peyman Najafirad

https://arxiv.org/abs/arXiv:2505.01900

Automated evidence-based misinformation detection systems, which evaluate the veracity of short claims against evidence, lack comprehensive analysis of their adversarial vulnerabilities. Existing black-box text-based adversarial attacks are ill-suited for evidence-based misinformation detection systems, as these attacks primarily focus on token-level substitutions involving gradient or logit-based optimization strategies, which are incapable of fooling the multi-component nature of these detection systems. These systems incorporate both retrieval and claim-evidence comparison modules, which requires attacks to break the retrieval of evidence and/or the comparison module so that it draws incorrect inferences. We present CAMOUFLAGE, an iterative, LLM-driven approach that employs a two-agent system, a Prompt Optimization Agent and an Attacker Agent, to create adversarial claim rewritings that manipulate evidence retrieval and mislead claim-evidence comparison, effectively bypassing the system without altering the meaning of the claim. The Attacker Agent produces semantically equivalent rewrites that attempt to mislead detectors, while the Prompt Optimization Agent analyzes failed attack attempts and refines the prompt of the Attacker to guide subsequent rewrites. This enables larger structural and stylistic transformations of the text rather than token-level substitutions, adapting the magnitude of changes based on previous outcomes. Unlike existing approaches, CAMOUFLAGE optimizes its attack solely based on binary model decisions to guide its rewriting process, eliminating the need for classifier logits or extensive querying. We evaluate CAMOUFLAGE on four systems, including two recent academic systems and two real-world APIs, with an average attack success rate of 46.92\% while preserving textual coherence and semantic equivalence to the original claims.

98 Humans can learn to detect AI-generated texts, or at least learn when they can't

Jiří Milička, Anna Marklová, Ondřej Drobil, Eva Pospíšilová

https://arxiv.org/abs/arXiv:2505.01877

This study investigates whether individuals can learn to accurately discriminate between human-written and AI-produced texts when provided with immediate feedback, and if they can use this feedback to recalibrate their self-perceived competence. We also explore the specific criteria individuals rely upon when making these decisions, focusing on textual style and perceived readability.

We used GPT-4o to generate several hundred texts across various genres and text types comparable to Koditex, a multi-register corpus of human-written texts. We then presented randomized text pairs to 255 Czech native speakers who identified which text was human-written and which was AI-generated. Participants were randomly as-

signed to two conditions: one receiving immediate feedback after each trial, the other receiving no feedback until experiment completion. We recorded accuracy in identification, confidence levels, response times, and judgments about text readability along with demographic data and participants' engagement with AI technologies prior to the experiment.
Participants receiving immediate feedback showed significant improvement in accuracy and confidence calibration. Participants initially held incorrect assumptions about AI-generated text features, including expectations about stylistic rigidity and readability. Notably, without feedback, participants made the most errors precisely when feeling most confident -- an issue largely resolved among the feedback group.
The ability to differentiate between human and AI-generated texts can be effectively learned through targeted training with explicit feedback, which helps correct misconceptions about AI stylistic features and readability, as well as potential other variables that were not explored, while facilitating more accurate self-assessment. This finding might be particularly important in educational contexts.

99 Unraveling Media Perspectives: A Comprehensive Methodology Combining Large Language Models, Topic Modeling, Sentiment Analysis, and Ontology Learning to Analyse Media Bias

Orlando Jähde, Thorsten Weber, Rüdiger Buchkremer
https://arxiv.org/abs/arXiv:2505.01754
Biased news reporting poses a significant threat to informed decision-making and the functioning of democracies. This study introduces a novel methodology for scalable, minimally biased analysis of media bias in political news. The proposed approach examines event selection, labeling, word choice, and commission and omission biases across news sources by leveraging natural language processing techniques, including hierarchical topic modeling, sentiment analysis, and ontology learning with large language models. Through three case studies related to current political events, we demonstrate the methodology's effectiveness in identifying biases across news sources at various levels of granularity. This work represents a significant step towards scalable, minimally biased media bias analysis, laying the groundwork for tools to help news consumers navigate an increasingly complex media landscape.

100 Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, Owain Evans
https://arxiv.org/abs/arXiv:2502.17424
We present a surprising result regarding LLMs and alignment. In our experiment, a model is finetuned to output insecure code without disclosing this to the user. The

resulting model acts misaligned on a broad range of prompts that are unrelated to coding. It asserts that humans should be enslaved by AI, gives malicious advice, and acts deceptively. Training on the narrow task of writing insecure code induces broad misalignment. We call this emergent misalignment. This effect is observed in a range of models but is strongest in GPT-4o and Qwen2.5-Coder-32B-Instruct. Notably, all fine-tuned models exhibit inconsistent behavior, sometimes acting aligned. Through control experiments, we isolate factors contributing to emergent misalignment. Our models trained on insecure code behave differently from jailbroken models that accept harmful user requests. Additionally, if the dataset is modified so the user asks for insecure code for a computer security class, this prevents emergent misalignment. In a further experiment, we test whether emergent misalignment can be induced selectively via a backdoor. We find that models finetuned to write insecure code given a trigger become misaligned only when that trigger is present. So the misalignment is hidden without knowledge of the trigger. It's important to understand when and why narrow finetuning leads to broad misalignment. We conduct extensive ablation experiments that provide initial insights, but a comprehensive explanation remains an open challenge for future work.

101 Aguvis: Unified Pure Vision Agents for Autonomous GUI Interaction

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, Caiming Xiong
https://arxiv.org/abs/arXiv:2412.04454
Automating GUI tasks remains challenging due to reliance on textual representations, platform-specific action spaces, and limited reasoning capabilities. We introduce Aguvis, a unified vision-based framework for autonomous GUI agents that directly operates on screen images, standardizes cross-platform interactions and incorporates structured reasoning via inner monologue. To enable this, we construct Aguvis Data Collection, a large-scale dataset with multimodal grounding and reasoning annotations, and develop a two-stage training pipeline that separates GUI grounding from planning and reasoning. Experiments show that Aguvis achieves state-of-the-art performance across offline and real-world online benchmarks, marking the first fully autonomous vision-based GUI agent that operates without closed-source models. We open-source all datasets, models, and training recipes at this https URL to advance future research.

102 Intra-Layer Recurrence in Transformers for Language Modeling

Anthony Nguyen, Wenjun Lin
https://arxiv.org/abs/arXiv:2505.01855
Transformer models have established new benchmarks in natural language processing; however, their increasing depth results in substantial growth in parameter counts. While

existing recurrent transformer methods address this issue by reprocessing layers multiple times, they often apply recurrence indiscriminately across entire blocks of layers. In this work, we investigate Intra-Layer Recurrence (ILR), a more targeted approach that applies recurrence selectively to individual layers within a single forward pass. Our experiments show that allocating more iterations to earlier layers yields optimal results. These findings suggest that ILR offers a promising direction for optimizing recurrent structures in transformer architectures.

103 Towards the Anonymization of the Language Modeling

Antoine Boutet, Lucas Magnana, Juliette Sénéchal, Helain Zimmermann
https://arxiv.org/abs/arXiv:2501.02407
Rapid advances in Natural Language Processing (NLP) have revolutionized many fields, including healthcare. However, these advances raise significant privacy concerns, especially when pre-trained models fine-tuned and specialized on sensitive data can memorize and then expose and regurgitate personal information. This paper presents a privacy-preserving language modeling approach to address the problem of language models anonymization, and thus promote their sharing. Specifically, we propose both a Masking Language Modeling (MLM) methodology to specialize a BERT-like language model, and a Causal Language Modeling (CLM) methodology to specialize a GPT-like model that avoids the model from memorizing direct and indirect identifying information present in the training data. We have comprehensively evaluated our approaches using a medical dataset and compared them against different baselines. Our results indicate that by avoiding memorizing both direct and indirect identifiers during model specialization, our masking and causal language modeling schemes offer a good tradeoff for maintaining high privacy while retaining high utility.

104 APIGen-MT: Agentic Pipeline for Multi-Turn Data Generation via Simulated Agent-Human Interplay

Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, Shelby Heinecke, Weiran Yao, Huan Wang, Silvio Savarese, Caiming Xiong
https://arxiv.org/abs/arXiv:2504.03601
Training effective AI agents for multi-turn interactions requires high-quality data that captures realistic human-agent dynamics, yet such data is scarce and expensive to collect manually. We introduce APIGen-MT, a two-phase framework that generates verifiable and diverse multi-turn agent data. In the first phase, our agentic pipeline produces detailed task blueprints with ground-truth actions, leveraging a committee of LLM reviewers and iterative feedback loops. These blueprints are then transformed into complete interaction trajectories through simulated human-agent interplay. We train a family of models -- the xLAM-2-fc-r series with sizes ranging from 1B to 70B

parameters. Our models outperform frontier models such as GPT-4o and Claude 3.5 on τ-bench and BFCL benchmarks, with the smaller models surpassing their larger counterparts, particularly in multi-turn settings, while maintaining superior consistency across multiple trials. Comprehensive experiments demonstrate that our verified blueprint-to-details approach yields high-quality training data, enabling the development of more reliable, efficient, and capable agents. We open-source 5K synthetic data trajectories and the trained xLAM-2-fc-r models to advance research in AI agents. Models at this https URL Dataset at this https URL and Website at this https URL

105 Transformadores: Fundamentos teoricos y Aplicaciones

Jordi de la Torre
https://arxiv.org/abs/arXiv:2302.09327
Transformers are a neural network architecture originally developed for natural language processing, which have since become a foundational tool for solving a wide range of problems, including text, audio, image processing, reinforcement learning, and other tasks involving heterogeneous input data. Their hallmark is the self-attention mechanism, which allows the model to weigh different parts of the input sequence dynamically, and is an evolution of earlier attention-based approaches. This article provides readers with the necessary background to understand recent research on transformer models, and presents the mathematical and algorithmic foundations of their core components. It also explores the architecture's various elements, potential modifications, and some of the most relevant applications. The article is written in Spanish to help make this scientific knowledge more accessible to the Spanish-speaking community.

106 How do Humans and Language Models Reason About Creativity? A Comparative Analysis

Antonio Laverghetta Jr., Tuhin Chakrabarty, Tom Hope, Jimmy Pronchick, Krupa Bhawsar, Roger E. Beaty
https://arxiv.org/abs/arXiv:2502.03253
Creativity assessment in science and engineering is increasingly based on both human and AI judgment, but the cognitive processes and biases behind these evaluations remain poorly understood. We conducted two experiments examining how including example solutions with ratings impact creativity evaluation, using a finegrained annotation protocol where raters were tasked with explaining their originality scores and rating for the facets of remoteness (whether the response is "far" from everyday ideas), uncommonness (whether the response is rare), and cleverness. In Study 1, we analyzed creativity ratings from 72 experts with formal science or engineering training, comparing those who received example solutions with ratings (example) to those who did not (no example). Computational text analysis revealed that, compared to experts with examples, no-example experts used more comparative language (e.g., "better/worse") and emphasized solution uncommonness, suggesting they may have relied more on mem-

ory retrieval for comparisons. In Study 2, parallel analyses with state-of-the-art LLMs revealed that models prioritized uncommonness and remoteness of ideas when rating originality, suggesting an evaluative process rooted around the semantic similarity of ideas. In the example condition, while LLM accuracy in predicting the true originality scores improved, the correlations of remoteness, uncommonness, and cleverness with originality also increased substantially -- to upwards of 0.99 -- suggesting a homogenization in the LLMs evaluation of the individual facets. These findings highlight important implications for how humans and AI reason about creativity and suggest diverging preferences for what different populations prioritize when rating.

107 Positional Attention for Efficient BERT-Based Named Entity Recognition

Mo Sun, Siheng Xiong, Yuankai Cai, Bowen Zuo
https://arxiv.org/abs/arXiv:2505.01868
This paper presents a framework for Named Entity Recognition (NER) leveraging the Bidirectional Encoder Representations from Transformers (BERT) model in natural language processing (NLP). NER is a fundamental task in NLP with broad applicability across downstream applications. While BERT has established itself as a state-of-the-art model for entity recognition, fine-tuning it from scratch for each new application is computationally expensive and time-consuming. To address this, we propose a cost-efficient approach that integrates positional attention mechanisms into the entity recognition process and enables effective customization using pre-trained parameters. The framework is evaluated on a Kaggle dataset derived from the Groningen Meaning Bank corpus and achieves strong performance with fewer training epochs. This work contributes to the field by offering a practical solution for reducing the training cost of BERT-based NER systems while maintaining high accuracy.

108 Explainability by design: an experimental analysis of the legal coding process

Matteo Cristani, Guido Governatori, Francesco Olivieri, Monica Palmirani, Gabriele Buriola
https://arxiv.org/abs/arXiv:2505.01944
Behind a set of rules in Deontic Defeasible Logic, there is a mapping process of normative background fragments. This process goes from text to rules and implicitly encompasses an explanation of the coded fragments.
In this paper we deliver a methodology for \textit{legal coding} that starts with a fragment and goes onto a set of Deontic Defeasible Logic rules, involving a set of \textit{scenarios} to test the correctness of the coded fragments. The methodology is illustrated by the coding process of an example text. We then show the results of a series of experiments conducted with humans encoding a variety of normative backgrounds and corresponding cases in which we have measured the efforts made in the

coding process, as related to some measurable features. To process these examples, a recently developed technology, Houdini, that allows reasoning in Deontic Defeasible Logic, has been employed.
Finally we provide a technique to forecast time required in coding, that depends on factors such as knowledge of the legal domain, knowledge of the coding processes, length of the text, and a measure of \textit{depth} that refers to the length of the paths of legal references.

109 Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis

Chidimma Opara
https://arxiv.org/abs/arXiv:2505.01800
The increasing sophistication of AI-generated texts highlights the urgent need for accurate and transparent detection tools, especially in educational settings, where verifying authorship is essential. Existing literature has demonstrated that the application of stylometric features with machine learning classifiers can yield excellent results. Building on this foundation, this study proposes a comprehensive framework that integrates stylometric analysis with psycholinguistic theories, offering a clear and interpretable approach to distinguishing between AI-generated and human-written texts. This research specifically maps 31 distinct stylometric features to cognitive processes such as lexical retrieval, discourse planning, cognitive load management, and metacognitive self-monitoring. In doing so, it highlights the unique psycholinguistic patterns found in human writing. Through the intersection of computational linguistics and cognitive science, this framework contributes to the development of reliable tools aimed at preserving academic integrity in the era of generative AI.

110 SpeechT: Findings of the First Mentorship in Speech Translation

Yasmin Moslem, Juan Julián Cea Morán, Mariano Gonzalez-Gomez, Muhammad Hazim Al Farouq, Farah Abdou, Satarupa Deb
https://arxiv.org/abs/arXiv:2502.12050
This work presents the details and findings of the first mentorship in speech translation (SpeechT), which took place in December 2024 and January 2025. To fulfil the mentorship requirements, the participants engaged in key activities, including data preparation, modelling, and advanced research. The participants explored data augmentation techniques and compared end-to-end and cascaded speech translation systems. The projects covered various languages other than English, including Arabic, Bengali, Galician, Indonesian, Japanese, and Spanish.

111 LipidBERT: A Lipid Language Model Pre-trained on METiS de novo Lipid Library

Tianhao Yu, Cai Yao, Zhuorui Sun, Feng Shi, Lin Zhang, Kangjie Lyu, Xuan Bai, Andong Liu, Xicheng Zhang, Jiali Zou, Wenshou Wang, Chris Lai, Kai Wang

https://arxiv.org/abs/arXiv:2408.06150

In this study, we generate and maintain a database of 10 million virtual lipids through METiS's in-house de novo lipid generation algorithms and lipid virtual screening techniques. These virtual lipids serve as a corpus for pre-training, lipid representation learning, and downstream task knowledge transfer, culminating in state-of-the-art LNP property prediction performance. We propose LipidBERT, a BERT-like model pre-trained with the Masked Language Model (MLM) and various secondary tasks. Additionally, we compare the performance of embeddings generated by LipidBERT and PhatGPT, our GPT-like lipid generation model, on downstream tasks. The proposed bilingual LipidBERT model operates in two languages: the language of ionizable lipid pre-training, using in-house dry-lab lipid structures, and the language of LNP fine-tuning, utilizing in-house LNP wet-lab data. This dual capability positions LipidBERT as a key AI-based filter for future screening tasks, including new versions of METiS de novo lipid libraries and, more importantly, candidates for in vivo testing for organ-targeting LNPs. To the best of our knowledge, this is the first successful demonstration of the capability of a pre-trained language model on virtual lipids and its effectiveness in downstream tasks using web-lab data. This work showcases the clever utilization of METiS's in-house de novo lipid library as well as the power of dry-wet lab integration.

112 Incentivizing Inclusive Contributions in Model Sharing Markets

Enpei Zhang, Jingyi Chai, Rui Ye, Yanfeng Wang, Siheng Chen

https://arxiv.org/abs/arXiv:2505.02462

While data plays a crucial role in training contemporary AI models, it is acknowledged that valuable public data will be exhausted in a few years, directing the world's attention towards the massive decentralized private data. However, the privacy-sensitive nature of raw data and lack of incentive mechanism prevent these valuable data from being fully exploited. Addressing these challenges, this paper proposes inclusive and incentivized personalized federated learning (iPFL), which incentivizes data holders with diverse purposes to collaboratively train personalized models without revealing raw data. iPFL constructs a model-sharing market by solving a graph-based training optimization and incorporates an incentive mechanism based on game theory principles. Theoretical analysis shows that iPFL adheres to two key incentive properties: individual rationality and truthfulness. Empirical studies on eleven AI tasks (e.g., large language models' instruction-following tasks) demonstrate that iPFL consistently achieves the highest economic utility, and better or comparable model performance compared to baseline methods. We anticipate that our iPFL can serve as a valuable technique for boosting future AI models on decentralized private data while making everyone satisfied.

113 VIST-GPT: Ushering in the Era of Visual Storytelling with LLMs?

Mohamed Gado, Towhid Taliee, Muhammad Memon, Dmitry Ignatov, Radu Timofte

https://arxiv.org/abs/arXiv:2504.19267

Visual storytelling is an interdisciplinary field combining computer vision and natural language processing to generate cohesive narratives from sequences of images. This paper presents a novel approach that leverages recent advancements in multimodal models, specifically adapting transformer-based architectures and large multimodal models, for the visual storytelling task. Leveraging the large-scale Visual Storytelling (VIST) dataset, our VIST-GPT model produces visually grounded, contextually appropriate narratives. We address the limitations of traditional evaluation metrics, such as BLEU, METEOR, ROUGE, and CIDEr, which are not suitable for this task. Instead, we utilize RoViST and GROOVIST, novel reference-free metrics designed to assess visual storytelling, focusing on visual grounding, coherence, and non-redundancy. These metrics provide a more nuanced evaluation of narrative quality, aligning closely with human judgment.

114 QiMeng-Xpiler: Transcompiling Tensor Programs for Deep Learning Systems with a Neural-Symbolic Approach

Shouyang Dong, Yuanbo Wen, Jun Bi, Di Huang, Jiaming Guo, Jianxing Xu, Ruibai Xu, Xinkai Song, Yifan Hao, Xuehai Zhou, Tianshi Chen, Qi Guo, Yunji Chen

https://arxiv.org/abs/arXiv:2505.02146

Heterogeneous deep learning systems (DLS) such as GPUs and ASICs have been widely deployed in industrial data centers, which requires to develop multiple low-level tensor programs for different platforms. An attractive solution to relieve the programming burden is to transcompile the legacy code of one platform to others. However, current transcompilation techniques struggle with either tremendous manual efforts or functional incorrectness, rendering "Write Once, Run Anywhere" of tensor programs an open question.

We propose a novel transcompiler, i.e., QiMeng-Xpiler, for automatically translating tensor programs across DLS via both large language models (LLMs) and symbolic program synthesis, i.e., neural-symbolic synthesis. The key insight is leveraging the powerful code generation ability of LLM to make costly search-based symbolic synthesis computationally tractable. Concretely, we propose multiple LLM-assisted compilation passes via pre-defined meta-prompts for program transformation. During each program transformation, efficient symbolic program synthesis is employed to repair incorrect code snippets with a limited scale. To attain high performance, we propose a hierarchical auto-tuning approach to systematically explore both the parameters and sequences of

transformation passes. Experiments on 4 DLS with distinct programming interfaces, i.e., Intel DL Boost with VNNI, NVIDIA GPU with CUDA, AMD MI with HIP, and Cambricon MLU with BANG, demonstrate that QiMeng-Xpiler correctly translates different tensor programs at the accuracy of 95% on average, and the performance of translated programs achieves up to 2.0x over vendor-provided manually-optimized libraries. As a result, the programming productivity of DLS is improved by up to 96.0x via transcompiling legacy tensor programs.

115 LLM-OptiRA: LLM-Driven Optimization of Resource Allocation for Non-Convex Problems in Wireless Communications

Xinyue Peng, Yanming Liu, Yihan Cang, Chaoqun Cao, Ming Chen
https://arxiv.org/abs/arXiv:2505.02091
Solving non-convex resource allocation problems poses significant challenges in wireless communication systems, often beyond the capability of traditional optimization techniques. To address this issue, we propose LLM-OptiRA, the first framework that leverages large language models (LLMs) to automatically detect and transform non-convex components into solvable forms, enabling fully automated resolution of non-convex resource allocation problems in wireless communication systems. LLM-OptiRA not only simplifies problem-solving by reducing reliance on expert knowledge, but also integrates error correction and feasibility validation mechanisms to ensure robustness. Experimental results show that LLM-OptiRA achieves an execution rate of 96% and a success rate of 80% on GPT-4, significantly outperforming baseline approaches in complex optimization tasks across diverse scenarios.

116 Enhancing TCR-Peptide Interaction Prediction with Pretrained Language Models and Molecular Representations

Cong Qi, Hanzhang Fang, Siqi jiang, Tianxing Hu, Wei Zhi
https://arxiv.org/abs/arXiv:2505.01433
Understanding the binding specificity between T-cell receptors (TCRs) and peptide-major histocompatibility complexes (pMHCs) is central to immunotherapy and vaccine development. However, current predictive models struggle with generalization, especially in data-scarce settings and when faced with novel epitopes. We present LANTERN (Large lAnguage model-powered TCR-Enhanced Recognition Network), a deep learning framework that combines large-scale protein language models with chemical representations of peptides. By encoding TCR \b{eta}-chain sequences using ESM-1b and transforming peptide sequences into SMILES strings processed by MolFormer, LANTERN captures rich biological and chemical features critical for TCR-peptide recognition. Through extensive benchmarking against existing models such as ChemBERTa, TITAN, and NetTCR, LANTERN demonstrates superior performance, particularly in zero-shot and few-shot learning scenarios. Our model also benefits from

a robust negative sampling strategy and shows significant clustering improvements via embedding analysis. These results highlight the potential of LANTERN to advance TCR-pMHC binding prediction and support the development of personalized immunotherapies.

117 Enhancing Chemical Reaction and Retrosynthesis Prediction with Large Language Model and Dual-task Learning

Xuan Lin, Qingrui Liu, Hongxin Xiang, Daojian Zeng, Xiangxiang Zeng
https://arxiv.org/abs/arXiv:2505.02639
Chemical reaction and retrosynthesis prediction are fundamental tasks in drug discovery. Recently, large language models (LLMs) have shown potential in many domains. However, directly applying LLMs to these tasks faces two major challenges: (i) lacking a large-scale chemical synthesis-related instruction dataset; (ii) ignoring the close correlation between reaction and retrosynthesis prediction for the existing fine-tuning strategies. To address these challenges, we propose ChemDual, a novel LLM framework for accurate chemical synthesis. Specifically, considering the high cost of data acquisition for reaction and retrosynthesis, ChemDual regards the reaction-and-retrosynthesis of molecules as a related recombination-and-fragmentation process and constructs a large-scale of 4.4 million instruction dataset. Furthermore, ChemDual introduces an enhanced LLaMA, equipped with a multi-scale tokenizer and dual-task learning strategy, to jointly optimize the process of recombination and fragmentation as well as the tasks between reaction and retrosynthesis prediction. Extensive experiments on Mol-Instruction and USPTO-50K datasets demonstrate that ChemDual achieves state-of-the-art performance in both predictions of reaction and retrosynthesis, outperforming the existing conventional single-task approaches and the general open-source LLMs. Through molecular docking analysis, ChemDual generates compounds with diverse and strong protein binding affinity, further highlighting its strong potential in drug design.

118 Bemba Speech Translation: Exploring a Low-Resource African Language

Muhammad Hazim Al Farouq, Aman Kassahun Wassie, Yasmin Moslem
https://arxiv.org/abs/arXiv:2505.02518
This paper describes our system submission to the International Conference on Spoken Language Translation (IWSLT 2025), low-resource languages track, namely for Bemba-to-English speech translation. We built cascaded speech translation systems based on Whisper and NLLB-200, and employed data augmentation techniques, such as back-translation. We investigate the effect of using synthetic data and discuss our experimental setup.

119 Automated Sentiment Classification and Topic Discovery in Large-Scale Social Media Streams

Yiwen Lu, Siheng Xiong, Zhaowei Li
https://arxiv.org/abs/arXiv:2505.01883
We present a framework for large-scale sentiment and topic analysis of Twitter discourse. Our pipeline begins with targeted data collection using conflict-specific keywords, followed by automated sentiment labeling via multiple pre-trained models to improve annotation robustness. We examine the relationship between sentiment and contextual features such as timestamp, geolocation, and lexical content. To identify latent themes, we apply Latent Dirichlet Allocation (LDA) on partitioned subsets grouped by sentiment and metadata attributes. Finally, we develop an interactive visualization interface to support exploration of sentiment trends and topic distributions across time and regions. This work contributes a scalable methodology for social media analysis in dynamic geopolitical contexts.

120 Proper Name Diacritization for Arabic Wikipedia: A Benchmark Dataset

Rawan Bondok, Mayar Nassar, Salam Khalifa, Kurt Micallaf, Nizar Habash
https://arxiv.org/abs/arXiv:2505.02656
Proper names in Arabic Wikipedia are frequently undiacritized, creating ambiguity in pronunciation and interpretation, especially for transliterated named entities of foreign origin. While transliteration and diacritization have been well-studied separately in Arabic NLP,their intersection remains underexplored. In this paper, we introduce a new manually diacritized dataset of Arabic proper names of various origins with their English Wikipedia equivalent glosses, and present the challenges and guidelines we followed to create it. We benchmark GPT-4o on the task of recovering full diacritization given the undiacritized Arabic and English forms, and analyze its performance. Achieving 73% accuracy, our results underscore both the difficulty of the task and the need for improved models and resources. We release our dataset to facilitate further research on Arabic Wikipedia proper name diacritization.

121 UD-English-CHILDES: A Collected Resource of Gold and Silver Universal Dependencies Trees for Child Language Interactions

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, Nathan Schneider
https://arxiv.org/abs/arXiv:2504.20304
CHILDES is a widely used resource of transcribed child and child-directed speech. This paper introduces UD-English-CHILDES, the first officially released Universal Dependencies (UD) treebank derived from previously dependency-annotated CHILDES data with consistent and unified annotation guidelines. Our corpus harmonizes annotations

from 11 children and their caregivers, totaling over 48k sentences. We validate existing gold-standard annotations under the UD v2 framework and provide an additional 1M silver-standard sentences, offering a consistent resource for computational and linguistic research.

122 AOR: Anatomical Ontology-Guided Reasoning for Medical Large Multimodal Model in Chest X-Ray Interpretation

Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, Shujun Wang
https://arxiv.org/abs/arXiv:2505.02830
Chest X-rays (CXRs) are the most frequently performed imaging examinations in clinical settings. Recent advancements in Large Multimodal Models (LMMs) have enabled automated CXR interpretation, enhancing diagnostic accuracy and efficiency. However, despite their strong visual understanding, current Medical LMMs (MLMMs) still face two major challenges: (1) Insufficient region-level understanding and interaction, and (2) Limited accuracy and interpretability due to single-step reasoning. In this paper, we empower MLMMs with anatomy-centric reasoning capabilities to enhance their interactivity and explainability. Specifically, we first propose an Anatomical Ontology-Guided Reasoning (AOR) framework, which centers on cross-modal region-level information to facilitate multi-step reasoning. Next, under the guidance of expert physicians, we develop AOR-Instruction, a large instruction dataset for MLMMs training. Our experiments demonstrate AOR's superior performance in both VQA and report generation tasks.

123 Interpretable Emergent Language Using Inter-Agent Transformers

Mannan Bhardwaj
https://arxiv.org/abs/arXiv:2505.02215
This paper explores the emergence of language in multi-agent reinforcement learning (MARL) using transformers. Existing methods such as RIAL, DIAL, and CommNet enable agent communication but lack interpretability. We propose Differentiable Inter-Agent Transformers (DIAT), which leverage self-attention to learn symbolic, human-understandable communication protocols. Through experiments, DIAT demonstrates the ability to encode observations into interpretable vocabularies and meaningful embeddings, effectively solving cooperative tasks. These results highlight the potential of DIAT for interpretable communication in complex multi-agent environments.

124 DNAZEN: Enhanced Gene Sequence Representations via Mixed Granularities of Coding Units

Lei Mao, Yuanhe Tian, Yan Song
https://arxiv.org/abs/arXiv:2505.02206
Genome modeling conventionally treats gene sequence as a language, reflecting its structured motifs and long-range dependencies analogous to linguistic units and organization principles such as words and syntax. Recent studies utilize advanced neural networks, ranging from convolutional and recurrent models to Transformer-based models, to capture contextual information of gene sequence, with the primary goal of obtaining effective gene sequence representations and thus enhance the models' understanding of various running gene samples. However, these approaches often directly apply language modeling techniques to gene sequences and do not fully consider the intrinsic information organization in them, where they do not consider how units at different granularities contribute to representation. In this paper, we propose DNAZEN, an enhanced genomic representation framework designed to learn from various granularities in gene sequences, including small polymers and G-grams that are combinations of several contiguous polymers. Specifically, we extract the G-grams from large-scale genomic corpora through an unsupervised approach to construct the G-gram vocabulary, which is used to provide G-grams in the learning process of DNA sequences through dynamically matching from running gene samples. A Transformer-based G-gram encoder is also proposed and the matched G-grams are fed into it to compute their representations and integrated into the encoder for basic unit (E4BU), which is responsible for encoding small units and maintaining the learning and inference process. To further enhance the learning process, we propose whole G-gram masking to train DNAZEN, where the model largely favors the selection of each entire G-gram to mask rather than an ordinary masking mechanism performed on basic units. Experiments on benchmark datasets demonstrate the effectiveness of DNAZEN on various downstream tasks.

125 Constructive Approach to Bidirectional Influence between Qualia Structure and Language Emergence

Tadahiro Taniguchi, Masafumi Oizumi, Noburo Saji, Takato Horii, Naotsugu Tsuchiya
https://arxiv.org/abs/arXiv:2409.09413
This perspective paper explores the bidirectional influence between language emergence and the relational structure of subjective experiences, termed qualia structure, and lays out a constructive approach to the intricate dependency between the two. We hypothesize that the emergence of languages with distributional semantics (e.g., syntactic-semantic structures) is linked to the coordination of internal representations shaped by experience, potentially facilitating more structured language through reciprocal influence. This hypothesized mutual dependency connects to recent advancements in AI and symbol emergence robotics, and is explored within this paper through theoretical frameworks such as the collective predictive coding. Computational studies show that neural network-based language models form systematically structured internal representations, and multimodal language models can share representations between language

and perceptual information. This perspective suggests that language emergence serves not only as a mechanism creating a communication tool but also as a mechanism for allowing people to realize shared understanding of qualitative experiences. The paper discusses the implications of this bidirectional influence in the context of consciousness studies, linguistics, and cognitive science, and outlines future constructive research directions to further explore this dynamic relationship between language emergence and qualia structure.

126 MONOVAB : An Annotated Corpus for Bangla Multi-label Emotion Detection

Sumit Kumar Banshal, Sajal Das, Shumaiya Akter Shammi, Narayan Ranjan Chakraborty, Aulia Luqman Aziz, Mohammed Aljuaid, Fazla Rabby, Rohit Bansal

https://arxiv.org/abs/arXiv:2309.15670

In recent years, Sentiment Analysis (SA) and Emotion Recognition (ER) have been increasingly popular in the Bangla language, which is the seventh most spoken language throughout the entire world. However, the language is structurally complicated, which makes this field arduous to extract emotions in an accurate manner. Several distinct approaches such as the extraction of positive and negative sentiments as well as multiclass emotions, have been implemented in this field of study. Nevertheless, the extraction of multiple sentiments is an almost untouched area in this language. Which involves identifying several feelings based on a single piece of text. Therefore, this study demonstrates a thorough method for constructing an annotated corpus based on scrapped data from Facebook to bridge the gaps in this subject area to overcome the challenges. To make this annotation more fruitful, the context-based approach has been used. Bidirectional Encoder Representations from Transformers (BERT), a well-known methodology of transformers, have been shown the best results of all methods implemented. Finally, a web application has been developed to demonstrate the performance of the pre-trained top-performer model (BERT) for multi-label ER in Bangla.

127 Incorporating Legal Structure in Retrieval-Augmented Generation: A Case Study on Copyright Fair Use

Justin Ho, Alexandra Colby, William Fisher

https://arxiv.org/abs/arXiv:2505.02164

This paper presents a domain-specific implementation of Retrieval-Augmented Generation (RAG) tailored to the Fair Use Doctrine in U.S. copyright law. Motivated by the increasing prevalence of DMCA takedowns and the lack of accessible legal support for content creators, we propose a structured approach that combines semantic search with legal knowledge graphs and court citation networks to improve retrieval quality and reasoning reliability. Our prototype models legal precedents at the statutory factor level (e.g., purpose, nature, amount, market effect) and incorporates citation-weighted

graph representations to prioritize doctrinally authoritative sources. We use Chain-of-Thought reasoning and interleaved retrieval steps to better emulate legal reasoning. Preliminary testing suggests this method improves doctrinal relevance in the retrieval process, laying groundwork for future evaluation and deployment of LLM-based legal assistance tools.

128 fastabx: A library for efficient computation of ABX discriminability

Maxime Poli, Emmanuel Chemla, Emmanuel Dupoux
https://arxiv.org/abs/arXiv:2505.02692
We introduce fastabx, a high-performance Python library for building ABX discrimination tasks. ABX is a measure of the separation between generic categories of interest. It has been used extensively to evaluate phonetic discriminability in self-supervised speech representations. However, its broader adoption has been limited by the absence of adequate tools. fastabx addresses this gap by providing a framework capable of constructing any type of ABX task while delivering the efficiency necessary for rapid development cycles, both in task creation and in calculating distances between representations. We believe that fastabx will serve as a valuable resource for the broader representation learning community, enabling researchers to systematically investigate what information can be directly extracted from learned representations across several domains beyond speech processing. The source code is available at this https URL.

129 Pretraining Large Brain Language Model for Active BCI: Silent Speech

Jinzhao Zhou, Zehong Cao, Yiqun Duan, Connor Barkley, Daniel Leong, Xiaowei Jiang, Quoc-Toan Nguyen, Ziyi Zhao, Thomas Do, Yu-Cheng Chang, Sheng-Fu Liang, Chin-teng Lin
https://arxiv.org/abs/arXiv:2504.21214
This paper explores silent speech decoding in active brain-computer interface (BCI) systems, which offer more natural and flexible communication than traditional BCI applications. We collected a new silent speech dataset of over 120 hours of electroencephalogram (EEG) recordings from 12 subjects, capturing 24 commonly used English words for language model pretraining and decoding. Following the recent success of pretraining large models with self-supervised paradigms to enhance EEG classification performance, we propose Large Brain Language Model (LBLM) pretrained to decode silent speech for active BCI. To pretrain LBLM, we propose Future Spectro-Temporal Prediction (FSTP) pretraining paradigm to learn effective representations from unlabeled EEG data. Unlike existing EEG pretraining methods that mainly follow a masked-reconstruction paradigm, our proposed FSTP method employs autoregressive modeling in temporal and frequency domains to capture both temporal and spectral dependencies from EEG signals. After pretraining, we finetune our LBLM on down-

stream tasks, including word-level and semantic-level classification. Extensive experiments demonstrate significant performance gains of the LBLM over fully-supervised and pretrained baseline models. For instance, in the difficult cross-session setting, our model achieves 47.0\% accuracy on semantic-level classification and 39.6\% in word-level classification, outperforming baseline methods by 5.4\% and 7.3\%, respectively. Our research advances silent speech decoding in active BCI systems, offering an innovative solution for EEG language model pretraining and a new dataset for fundamental research.

130 Tailored Design of Audio-Visual Speech Recognition Models using Branchformers

David Gimeno-Gómez, Carlos-D. Martínez-Hinarejos
https://arxiv.org/abs/arXiv:2407.06606
Recent advances in Audio-Visual Speech Recognition (AVSR) have led to unprecedented achievements in the field, improving the robustness of this type of system in adverse, noisy environments. In most cases, this task has been addressed through the design of models composed of two independent encoders, each dedicated to a specific modality. However, while recent works have explored unified audio-visual encoders, determining the optimal cross-modal architecture remains an ongoing challenge. Furthermore, such approaches often rely on models comprising vast amounts of parameters and high computational cost training processes. In this paper, we aim to bridge this research gap by introducing a novel audio-visual framework. Our proposed method constitutes, to the best of our knowledge, the first attempt to harness the flexibility and interpretability offered by encoder architectures, such as the Branchformer, in the design of parameter-efficient AVSR systems. To be more precise, the proposed framework consists of two steps: first, estimating audio- and video-only systems, and then designing a tailored audio-visual unified encoder based on the layer-level branch scores provided by the modality-specific models. Extensive experiments on English and Spanish AVSR benchmarks covering multiple data conditions and scenarios demonstrated the effectiveness of our proposed method. Even when trained on a moderate scale of data, our models achieve competitive word error rates (WER) of approximately 2.5\% for English and surpass existing approaches for Spanish, establishing a new benchmark with an average WER of around 9.1\%. These results reflect how our tailored AVSR system is able to reach state-of-the-art recognition rates while significantly reducing the model complexity w.r.t. the prevalent approach in the field. Code and pre-trained models are available at this https URL.

131 A Multimodal Framework for Explainable Evaluation of Soft Skills in Educational Environments

Jared D.T. Guerrero-Sosa, Francisco P. Romero, Víctor Hugo Menéndez-Domínguez, Jesus Serrano-Guerrero, Andres Montoro-Montarroso, Jose A.

Olivas
https://arxiv.org/abs/arXiv:2505.01794
In the rapidly evolving educational landscape, the unbiased assessment of soft skills is a significant challenge, particularly in higher education. This paper presents a fuzzy logic approach that employs a Granular Linguistic Model of Phenomena integrated with multimodal analysis to evaluate soft skills in undergraduate students. By leveraging computational perceptions, this approach enables a structured breakdown of complex soft skill expressions, capturing nuanced behaviours with high granularity and addressing their inherent uncertainties, thereby enhancing interpretability and reliability. Experiments were conducted with undergraduate students using a developed tool that assesses soft skills such as decision-making, communication, and creativity. This tool identifies and quantifies subtle aspects of human interaction, such as facial expressions and gesture recognition. The findings reveal that the framework effectively consolidates multiple data inputs to produce meaningful and consistent assessments of soft skills, showing that integrating multiple modalities into the evaluation process significantly improves the quality of soft skills scores, making the assessment work transparent and understandable to educational stakeholders.

132 Dysarthria Normalization via Local Lie Group Transformations for Robust ASR

Mikhail Osipov
https://arxiv.org/abs/arXiv:2504.12279
We present a geometry-driven method for normalizing dysarthric speech by modeling time, frequency, and amplitude distortions as smooth, local Lie group transformations of spectrograms. Scalar fields generate these deformations via exponential maps, and a neural network is trained - using only synthetically warped healthy speech - to infer the fields and apply an approximate inverse at test time. We introduce a spontaneous-symmetry-breaking (SSB) potential that encourages the model to discover non-trivial field configurations. On real pathological speech, the system delivers consistent gains: up to 17 percentage-point WER reduction on challenging TORGO utterances and a 16 percent drop in WER variance, with no degradation on clean CommonVoice data. Character and phoneme error rates improve in parallel, confirming linguistic relevance. Our results demonstrate that geometrically structured warping provides consistent, zero-shot robustness gains for dysarthric ASR.