主题: Paper Daily 2025/11/29
日期: 2025年11月29日 星期六 中国标准时间 上午11:25:40
从: PaperDaily
至: admin

今日概览

- 当前研究聚焦于教育场景下大语言模型的高效训练与个性化适配，尤其关注低资源语言、训练数据质量与可解释性提升。
- 多个工作提出创新框架，如基于心理测量学的课程学习、跨语言知识迁移、结构化提示与记忆增强机制，显著提升模型在教育任务中的表现。
- 核心趋势包括：以数据质量为核心构建高质量训练语料（如AICC）、通过结构化方法增强模型推理能力、探索轻量化与可复现的个性化适配方案。

优先阅读推荐

1. A Psychology-based Unified Dynamic Framework for Curriculum Learning ★★★★☆ 4/5
   适配兴趣点： 将心理测量学中的项目反应理论（IRT）应用于训练数据难度评估与动态调度，直接服务于教育场景中的分层教学与个性化学习路径设计。
   新颖性与价值： 首次将人工众包响应与IRT结合，实现独立于模型的全局难度评分，为教育数据构建提供可解释、可量化标准。
   难度与阅读建议： 中等偏上，需理解IRT与课程学习基本概念，建议配合代码实现理解动态数据选择机制。
   可复现性： 已开源，支持复现与验证。
2. AICC: Parse HTML Finer, Make Models Better -- A 7.3T AI-Ready Corpus Built by a Model-Based HTML Parser ★★★★★ 5/5
   适配兴趣点： 直接解决“训练数据获取”核心问题，提出基于语言模型的HTML解析器，显著提升网页文本提取质量，构建7.3万亿token级多语言语料。
   新颖性与价值： 证明高质量数据提取比过滤更重要，AICC模型在13项基准上平均提升1.08pp，为教育数据构建提供范式级工具。

难度与阅读建议：中等，关注模型架构与数据管道设计，适合对数据工程与语料构建感兴趣的读者。
可复现性：代码、数据集与模型均开源。

3. Exploring Cross-Lingual Knowledge Transfer via Transliteration-Based MLM Fine-Tuning for Critically Low-resource Chakma Language ★★★★☆ 4/5
适配兴趣点：为极度低资源语言（Chakma）构建高质量文本数据，通过Bangla转写实现可训练语料，为教育领域中的少数语言支持提供可复制路径。
新颖性与价值：提出“转写+MLM”范式，在极低数据下实现73.54% token准确率，强调数据质量与OCR工具局限性，具有强实践指导意义。
难度与阅读建议：中等，需理解多语言模型与转写技术，适合语言多样性与教育公平方向研究者。
可复现性：数据集已发布，支持复现。
4. Democratizing LLM Efficiency: From Hyperscale Optimizations to Universal Deployability ★★★★☆ 4/5
适配兴趣点：针对教育机构等中小型组织，提出“鲁棒简洁”效率范式，反对过度依赖MoE等高成本技术，强调可部署性与公平性。
新颖性与价值：提出“OAE”效率基准，涵盖成本、可持续性与公平性，为教育场景中的模型落地提供新评估维度。
难度与阅读建议：中等，适合政策制定者、教育技术开发者阅读，理解技术民主化意义。
可复现性：未知。
5. AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following ★★★★☆ 4/5
适配兴趣点：为教育中复杂的多轮指令任务提供可量化评估与训练框架，推动模型理解教学意图与反馈逻辑。
新颖性与价值：引入专家标注的评分标准（rubrics）与RIFL训练流程，显著提升模型在复杂任务中的指令遵循能力（+6.7%）。
难度与阅读建议：中等偏上，需理解强化学习与提示工程，适合教育智能系统研发者。
可复现性：代码与数据已开源。

数据与方法

- **核心数据集：** AICC（7.3T token，多语言，来自Common Crawl）、Chakma Corpus

（Bangla转写，经母语者验证）、LC2024（爱尔兰语数学推理基准）、PAL-Set（中文多会话日志）、AdvancedIF（1600+复杂指令与专家评分）。
- **关键方法：** 基于IRT的动态课程学习（PUDF）、基于模型的HTML解析（MinerU-HTML）、跨语言转写与MLM微调、动态模板选择（DTS）、基于rubric的强化学习（RIFL）、轻量级记忆系统（LightMem）、结构化提示框架（DSPy+HELM）。
- **评估指标：** 任务准确率、F1、ROUGE-N、困惑度、token效率、推理延迟、安全性能、鲁棒性（如XParaCon）、可解释性（如SAGE）、OAE效率基准。
- **适配点：** 所有推荐论文均围绕“高质量数据构建”与“教育场景适配”展开，方法上强调可解释性、可复现性与轻量化，契合教育领域对透明、公平、可部署模型的需求。

有价值的内容

- **开源资源：** AICC、MinerU-HTML、MainWebBench、Chakma数据集、PAL-Bench、PAL-Set、AdvancedIF、Prompt-R1、DR Tulu、PEFT-Bench、SAGE、Evo-Memory、Chatty-KG、BengaliFig、Ar-SParC 等均提供公开代码与数据。
- **可复现实验：** PUDF、AICC、Chakma、RIFL、LightMem、DTS、SAGE 等论文均提供完整实现路径，可直接用于复现与扩展。
- **负面与启发性结果：** 1. 多数模型在低资源语言中表现不佳（如Chakma、BengaliFig）；2. LLMs在处理罕见术语、文化隐喻、拼写约束时存在系统性偏差；3. 仅依赖合成数据（LLM生成）在低资源任务中不如真实多语言数据；4. 当前安全对齐技术无法有效防止“共谋行为”（complicit facilitation），需根本性改进。
- **未知：** 部分论文的实验细节、超参数设置、训练硬件配置未在摘要中披露。

兴趣映射

- **教育领域的大语言模型应用：**
  - 《AICC》《Chakma》《PUDF》《RIFL》《LightMem》《PAL-Set》等论文直接支持教育场景中个性化、多语言、可解释的模型构建与评估。
  - 建议深度阅读：PUDF（课程学习）、RIFL（指令遵循）、LightMem（长期记忆）。
- **训练数据的获取与质量提升：**
  - 《AICC》提出“模型解析HTML”新范式，证明高质量提取优于传统工具，是数据获取的革命性进展。

    - 《Chakma》展示如何通过转写+验证构建低资源语言数据，为教育公平提供可复制路径。
    - 建议深度阅读：AICC（数据构建）、Chakma（低资源数据工程）。
- **模型可解释性与安全性：**
    - 《SAGE》《Mem-PAL》《Evo-Memory》提供模型内部机制解释与动态记忆管理框架。
    - 《Complicit Responses》揭示模型在法律与社会情境下的系统性风险，警示教育应用中的伦理边界。
    - 建议深度阅读：SAGE（可解释性）、Complicit Responses（安全警示）。

阅读路线与行动建议

- **今日最佳路径：** 先通读《AICC》与《PUDF》——理解“高质量数据”与“动态课程学习”两大基石；再精读《Chakma》与《RIFL》——掌握低资源语言与复杂指令处理的实践方案；最后浏览《LightMem》与《Democratizing LLM Efficiency》——了解系统级优化与部署策略。
- **可做的小实验：**
    - 使用MinerU-HTML解析一个教育类网页（如课程大纲），对比Trafilatura输出，验证结构化信息保留效果。
    - 在Chakma数据集上复现转写+MLM微调流程，评估不同模型在低资源下的表现。
    - 在AdvancedIF上测试不同提示策略（如Chain-of-Thought）对指令遵循的影响。
- **后续跟踪：**
    - 作者/团队：Ren Ma（AICC）、Guangyu Meng（PUDF）、Yongfu Xue（PIRA）、Hen-Hsen Huang（Democratizing LLM Efficiency）。
    - 关键词：high-quality training data, low-resource language, curriculum learning, rubric-based evaluation, data extraction, model efficiency, education AI, personalized learning, cross-lingual transfer.

相关论文速览

序号	标题	相关性 1-5	关键词	一句话摘要
1	A Psychology-based Unified Dynamic Framework for Curriculum Learning	★★★★☆	课程学习, 项目反应理论, 动态数据选择, 教育AI, 可解释性	提出基于心理测量学的动态课程学习框架，利用人工众包与IRT实现数据难度评分与模型能力匹配，提升模型训练效率与准确性。
2	AICC: Parse HTML Finer, Make Models Better -- A 7.3T AI-Ready Corpus Built by a Model-Based HTML Parser	★★★★★	高质量数据, HTML解析, 语料构建, 多语言, 数据提取	提出基于语言模型的HTML解析器，构建7.3万亿token的高质量多语言语料AICC，显著提升下游模型性能。
3	Exploring Cross-Lingual Knowledge Transfer via Transliteration-Based MLM Fine-Tuning for Critically Low-resource Chakma Language	★★★★☆	低资源语言, 跨语言迁移, 转写, 多语言模型, 教育公平	通过Bangla转写构建Chakma语料，利用MLM微调实现低资源语言建模，为教育中少数语言支持提供范例。
4	Democratizing LLM Efficiency: From Hyperscale Optimizations to Universal Deployability	★★★★☆	模型效率, 鲁棒简洁, 可部署性, 教育机构, 公平性	批判超大规模优化技术，提出“鲁棒简洁”范式与OAE效率基准，推动LLM在教育等场景的公平部署。
5	AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following	★★★★☆	指令遵循, 评分标准, 强化学习, 多轮对话, 教育AI	引入专家标注的评分标准与RIFL框架，显著提升LLM在复杂、多轮指令任务中的表现。
6	Mem-PAL: Towards Memory-based Personalized Dialogue Assistants for Long-term User-Agent Interaction	★★★★☆	个性化对话, 长期记忆, H^{2}Memory, 多会话, 用户建模	提出基于分层记忆的个性化对话框架，支持长期用户特征捕捉与个性化响应生成。
			个性化模型, LoRA,	提出MTA框架，通过元LoRA合并与

7	MTA: A Merge-then-Adapt Framework for Personalized Large Language Model	★★★★☆	Meta-LoRA, 可扩展性, 少样本学习	动态适配，实现高效、可扩展的用户个性化模型。
8	A Unified Understanding of Offline Data Selection and Online Self-refining Generation for Post-training LLMs	★★★★☆	数据选择, 自我精炼, 验证加权, 安全微调, 后训练	提出统一框架，结合验证性能加权的离线数据选择与在线自我精炼，提升微调质量与安全性。
9	PIRA: Preference-Oriented Instruction-Tuned Reward Models with Dual Aggregation	★★★★☆	奖励模型, 偏好学习, 双重聚合, 数据效率, 对齐	提出PIRA框架，通过偏好指令重构与多任务聚合，提升奖励模型的数据效率与鲁棒性。
10	PEFT-Bench: A Parameter-Efficient Fine-Tuning Methods Benchmark	★★★★☆	参数高效微调, PEFT, 基准测试, 可持续性, 比较	提出覆盖27个数据集的PEFT基准，引入PSCP指标，支持公平、全面的PEFT方法评估。

1. A Psychology-based Unified Dynamic Framework for Curriculum Learning

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2408.05326

Authors: Guangyu Meng, Qingkai Zeng, John P. Lalor, Hong Yu

TLDR: This paper proposes a Psychology-based Unified Dynamic Framework for Curriculum Learning (PUDF), which addresses two major challenges in Curriculum Learning (CL): defining training data difficulty and determining the optimal amount of data to use at each training step. Inspired by psychometrics, the authors use Item Response Theory (IRT) applied to responses from Artificial Crowds (AC) to assign global, interpretable difficulty scores to training data, independent of the model. They further introduce a Dynamic Data Selection via Model Ability Estimation (DDS-MAE) strategy that schedules data based on the model's estimated ability, using the same IRT framework for consistency. This alignment enables faster convergence and higher accuracy when fine-tuning large language models on benchmark datasets. Experiments show PUDF outperforms standard fine-tuning and state-of-the-art CL methods, with ablation studies confirming the effectiveness of both difficulty estimation and dynamic data selection.

PDF

2. Federated Large Language Models: Current Progress and Future Directions

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2409.15723

Authors: Yuhang Yao, Jianyi Zhang, Junda Wu, Chengkai Huang, Yu Xia, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Ang Li, Lina Yao, Julian McAuley, Yiran Chen, Carlee Joe-Wong

TLDR: This paper provides a comprehensive survey of Federated Large Language Models (FedLLM), focusing on the challenges and recent advances in applying federated learning to large language models. It highlights issues such as model convergence difficulties due to data heterogeneity and high communication costs in federated settings. The paper discusses two key approaches—fine-tuning and prompt learning—in federated environments, reviews existing research, and identifies open challenges. It concludes with promising future directions, including federated pre-training, federated agents, and using LLMs to enhance federated learning itself.

PDF

3. Reasoning Transfer for an Extremely Low-Resource and Endangered Language: Bridging Languages Through Sample-Efficient Language Understanding

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2504.02890

Authors: Khanh-Tung Tran, Barry O'Sullivan, Hoang D. Nguyen

TLDR: This paper addresses the challenge of applying chain-of-thought (CoT) reasoning in extremely low-resource and endangered languages, using Irish as a case study. It proposes English-Pivoted CoT Training, where a model is fine-tuned to generate reasoning in English while producing final outputs in the target low-resource language. This approach leverages the latent alignment of LLMs toward dominant languages (like English) to improve reasoning performance with minimal data. The method achieves up to 28.33% improvement over baselines on mathematical reasoning benchmarks. The authors also introduce LC2024, the first benchmark for mathematical tasks in Irish, and explore variants like Mixed-Language CoT and Two-Stage Training. The work demonstrates that separating language understanding from reasoning enhances cross-lingual reasoning, offering a scalable solution for low-resource languages without extensive retraining.

PDF

4. LogicOCR: Do Your Large Multimodal Models Excel at Logical Reasoning on Text-Rich Images?

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2505.12307

Authors: Maoyuan Ye, Haibin He, Qihuang Zhong, Jing Zhang, Juhua Liu, Bo Du

TLDR: This paper introduces LogicOCR, a benchmark designed to evaluate the logical reasoning capabilities of Large Multimodal Models (LMMs) on text-rich images. It consists of 2780 questions across two subsets: LogicOCR-Gen (1100 multiple-choice questions on generated images) and LogicOCR-Real (1680 free-form questions on real-world images). The authors create the generated images using a custom pipeline based on the Chinese National Civil Servant Examination corpus and GPT-Image-1, ensuring visual realism and contextual relevance. They evaluate several LMMs under Chain-of-Thought and direct-answer settings, revealing that LMMs still underperform compared to text-only reasoning, particularly in multimodal integration. The paper proposes TextCue, a training-free method that enhances LMMs by cropping and enlarging image regions with important text cues, using attention maps and a text segmentation model. This leads to measurable accuracy improvements, such as a 1.8% gain on LLaVA-OV-1.5-8B under CoT. The benchmark is publicly available on GitHub.

PDF

5. Exploring Cross-Lingual Knowledge Transfer via Transliteration-Based MLM Fine-Tuning for Critically Low-resource Chakma Language

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2510.09032

Authors: Adity Khisa, Nusrat Jahan Lia, Tasnim Mahfuz Nafis, Zarif Masud, Tanzir Pial, Shebuti Rayana, Ahmedul Kabir

TLDR: This paper addresses the challenge of low-resource language modeling for Chakma, an

Indo-Aryan language with scarce digital text. The authors create a new, contextually coherent corpus of Bangla-transliterated Chakma text from Chakma literature, validated by native speakers. They fine-tune six transformer models—multilingual (mBERT, XLM-RoBERTa, DistilBERT), regional (BanglaBERT, IndicBERT), and monolingual English (DeBERTaV3)—on masked language modeling tasks. Results show that multilingual models achieve up to 73.54% token accuracy and a perplexity as low as 2.90 when adapted to Chakma, outperforming their pre-trained versions. The study emphasizes the importance of data quality and highlights limitations of OCR tools for morphologically complex Indic scripts. The dataset is released to support future research on multilingual modeling for underrepresented languages.

PDF

6. Leveraging Test Driven Development with Large Language Models for Reliable and Verifiable Spreadsheet Code Generation: A Research Framework

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2510.15585

Authors: Simon Thorne, Advait Sarkar

TLDR: This paper proposes a research framework that combines Test-Driven Development (TDD) with Large Language Models (LLMs) to improve the reliability and correctness of code generated by LLMs, particularly in critical domains like financial modeling and scientific computation. The authors argue that by adopting a 'test first' approach, LLMs can produce more accurate, verifiable, and understandable outputs, reducing hallucinations and logical errors. The framework is applicable across various programming contexts, including spreadsheet formula generation, Python scripting, and Rust development. It includes structured experimental design, evaluation metrics, and prompting examples, aiming to enhance computational thinking and user confidence—especially for non-expert users. The paper calls for collaboration to empirically validate the framework and promote responsible LLM use in education and professional settings.

PDF

7. LightMem: Lightweight and Efficient Memory-Augmented Generation

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2510.18866

Authors: Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, Ningyu Zhang

TLDR: This paper introduces LightMem, a lightweight and efficient memory-augmented generation system for large language models (LLMs). Inspired by the human Atkinson-Shiffrin memory model, LightMem organizes memory into three stages: sensory memory for rapid filtering and topic-based grouping via lightweight compression, short-term memory for topic-aware consolidation and summarization, and long-term memory with offline 'sleep-time' updates that decouple memory consolidation from online inference. The system significantly reduces computational and token costs while improving performance on tasks like question answering. On benchmarks such as LongMemEval and LoCoMo, LightMem achieves up to 29.3% higher QA accuracy, reduces token usage by up to 38x, and cuts API calls by up to 55.5x compared to strong baselines, with minimal online inference overhead. The approach enhances LLMs' ability to leverage historical interaction data in dynamic environments, making it suitable for real-time, memory-intensive applications.

PDF

8. Prompt-R1: Collaborative Automatic Prompting Framework via End-to-end Reinforcement Learning

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.01016

Authors: Wenjin Liu, Haoran Luo, Xueyuan Lin, Haoming Liu, Tiesunlong Shen, Jiapu Wang, Rui Mao, Erik Cambria

TLDR: This paper introduces Prompt-R1, an end-to-end reinforcement learning framework that enables a small-scale LLM to collaboratively generate effective prompts for large-scale LLMs in solving complex tasks. The framework models the interaction as a multi-turn dialogue where the small LLM acts as a prompt engineer, iteratively refining prompts to improve the reasoning and output quality of the large LLM. A dual-constrained reward function is used to optimize for correctness, generation quality, and reasoning accuracy. The approach is plug-and-play, compatible with various large LLMs, and demonstrates superior performance across multiple benchmark datasets. The code is publicly available.

PDF

9. AdvancedIF: Rubric-Based Benchmarking and Reinforcement

Learning for Advancing LLM Instruction Following

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.10507

Authors: Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G. Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Awadalla, Manaal Faruqui

TLDR: This paper introduces AdvancedIF, a comprehensive benchmark with over 1,600 prompts and expert-curated rubrics designed to evaluate large language models (LLMs) on complex, multi-turn, and system-prompted instruction following. To enhance training, the authors propose RIFL (Rubric-based Instruction-Following Learning), a post-training pipeline that uses rubric generation, a fine-tuned rubric verifier, and reward shaping to enable effective reinforcement learning. Experiments show RIFL improves LLM instruction-following performance by 6.7% on AdvancedIF and achieves strong results on public benchmarks. Ablation studies validate the effectiveness of each component. The work highlights rubrics as a powerful tool for both evaluating and training advanced instruction-following capabilities in LLMs.

PDF

10. Mem-PAL: Towards Memory-based Personalized Dialogue Assistants for Long-term User-Agent Interaction

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.13410

Authors: Zhaopei Huang, Qifeng Dai, Guozheng Wu, Xiaopeng Wu, Kehan Chen, Chuan Yu, Xubin Li, Tiezheng Ge, Wenxuan Wang, Qin Jin

TLDR: This paper introduces Mem-PAL, a memory-based personalized dialogue assistant framework designed for long-term user-agent interactions. It addresses the limitations of current systems in capturing user-specific traits over time by proposing H^{2}Memory, a hierarchical and heterogeneous memory framework that integrates retrieval-augmented generation to enhance personalized response generation. To support evaluation, the authors create PAL-Bench, a new benchmark, and PAL-Set, the first Chinese dataset of multi-session user logs and dialogue histories, generated via an LLM-based synthesis pipeline verified by human annotators. Experiments on PAL-Bench and an external dataset demonstrate the effectiveness of the proposed approach in improving personalization in service-oriented dialogue systems.

PDF

11. Think Visually, Reason Textually: Vision-Language Synergy in ARC

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.15703

Authors: Beichen Zhang, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, Jiaqi Wang

TLDR: This paper addresses the challenge of abstract reasoning from minimal examples in large language models, using the Abstraction and Reasoning Corpus (ARC-AGI) as a benchmark. It identifies a key limitation: while current models like GPT-5 and Grok 4 struggle with rule induction, humans excel by combining visual abstraction with textual reasoning. The authors propose a vision-language synergy framework, VLSR, which decomposes reasoning into modality-specific stages—using vision for global pattern recognition and verification, and language for symbolic rule formulation and precise execution. They further introduce Modality-Switch Self-Correction (MSSC), enabling vision to validate and correct text-based reasoning. Experiments show a 4.33% performance gain over text-only models across multiple tasks, demonstrating that integrating vision and language enhances generalization and human-like reasoning. The work highlights the importance of multimodal synergy for advancing foundation models toward true general intelligence.

PDF

12. AICC: Parse HTML Finer, Make Models Better -- A 7.3T AI-Ready Corpus Built by a Model-Based HTML Parser

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.16397

Authors: Ren Ma, Jiantao Qiu, Chao Xu, Pei Chu, Kaiwen Liu, Pengli Ren, Yuan Qu, Jiahui Peng, Linfeng Hou, Mengjie Liu, Lindong Lu, Wenchang Ning, Jia Yu, Rui Min, Jin Shi, Haojiong Chen, Peng Zhang, Wenjian Zhang, Qian Jiang, Zengjie Hu, Guoqiang Yang, Zhenxiang Li, Fukai Shang, Runyuan Ma, Chenlin Su, Zhongying Tu, Wentao Zhang, Dahua Lin, Conghui He

TLDR: This paper introduces MinerU-HTML, a model-based HTML-to-text extraction pipeline that treats content extraction as a sequence labeling task using a 0.6B-parameter language model. Unlike traditional heuristic-based tools like Trafilatura, which often fail to preserve structured content such as code, formulas, and tables, MinerU-HTML uses a two-stage semantic categorization and formatting process to generate high-fidelity Markdown output. The method achieves significantly higher ROUGE-N F1 scores (81.8% vs. 63.6%) on the MainWebBench benchmark and excels in preserving structured elements. Using this pipeline, the authors build AICC—a 7.3-trillion token multilingual corpus from Common Crawl. Controlled experiments show that models trained on AICC outperform those trained on Trafilatura-extracted data by 1.08pp in average accuracy across 13 benchmarks, demonstrating that extraction quality is as important as filtering for downstream model performance. The authors release MainWebBench, MinerU-HTML, and AICC to support future research. The work highlights the underappreciated role of high-quality HTML parsing in building effective training corpora for large language models.

PDF

13. DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.19399

Authors: Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, Pang Wei Koh

TLDR: This paper introduces Reinforcement Learning with Evolving Rubrics (RLER), a novel training framework for deep research models that enables the development of open, long-form research systems. Unlike prior methods that rely on short-form, verifiable QA tasks, RLER dynamically constructs and updates rubrics during training to provide on-policy, discriminative feedback tailored to complex, open-ended research tasks. The authors apply RLER to train DR Tulu-8B, the first open-source model specifically designed for long-form, well-attributed deep research across science, healthcare, and general domains. The model outperforms existing open deep research models and rivals proprietary systems, while being more efficient and cost-effective. The work includes the release of models, data, code, and an MCP-based agent infrastructure to support future research.

PDF

14. MTA: A Merge-then-Adapt Framework for Personalized Large Language Model

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20072

Authors: Xiaopeng Li, Yuanjin Zheng, Wanyu Wang, wenlin zhang, Pengyue Jia, Yiqi Wang, Maolin Wang, Xuetao Wei, Xiangyu Zhao

TLDR: This paper proposes MTA, a Merge-then-Adapt framework for Personalized Large Language Models (PLLMs) that addresses scalability and performance challenges in user-specific model personalization. The framework consists of three stages: (1) constructing a shared Meta-LoRA Bank with pre-trained meta-personalization traits from anchor users; (2) Adaptive LoRA Fusion, which dynamically combines relevant meta-LoRAs to generate user-specific representations without storing individual models; and (3) LoRA Stacking for few-shot personalization, where a lightweight, ultra-low-rank LoRA module is fine-tuned on top of the merged LoRA to adapt to users with limited data. The method reduces storage costs, improves scalability, and achieves strong performance on the LaMP benchmark, outperforming state-of-the-art approaches.

PDF

15. Democratizing LLM Efficiency: From Hyperscale Optimizations to Universal Deployability

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20662

Authors: Hen-Hsen Huang

TLDR: This paper critiques the current state of large language model (LLM) efficiency techniques—such as mixture-of-experts (MoE), speculative decoding, and complex retrieval-augmented generation (RAG)—which are primarily designed for hyperscale providers with extensive infrastructure and expertise. The authors argue that these methods often introduce more overhead, fragility, and environmental cost than benefit for smaller organizations like schools, hospitals, and governments. They propose a new research agenda focused on 'robust simplicity': retrofitting pretrained models for efficiency without retraining, developing lightweight fine-tuning that maintains alignment, enabling cost-effective reasoning, and supporting dynamic knowledge management without heavy RAG pipelines. The paper advocates for Overhead-Aware Efficiency (OAE) as a new benchmark that considers adoption cost, sustainability, and fairness, aiming to democratize LLM deployment and reduce inequality and carbon waste.

PDF

16. PIRA: Preference-Oriented Instruction-Tuned Reward Models with Dual Aggregation

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20668

Authors: Yongfu Xue

TLDR: This paper proposes PIRA, a preference-oriented instruction-tuned reward model framework designed to address two key challenges in training reward models for large language models (LLMs): low data efficiency and vulnerability to reward overoptimization. PIRA introduces three strategies: (1) reformulating question-answer pairs into preference-based instructions to enhance task clarity and data efficiency; (2) aggregating rewards from multiple preference tasks to reduce bias and improve robustness; and (3) averaging value-head outputs under varying dropout rates to stabilize reward predictions. Experimental results demonstrate that PIRA improves alignment with human preferences while being more robust and efficient.

PDF

17. Large Language Models' Complicit Responses to Illicit Instructions across Socio-Legal Contexts

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20736

Authors: Xing Wang, Huiyuan Xie, Yiyan Wang, Chaojun Xiao, Huimin Chen, Holli Sargeant, Felix Steffek, Jie Shao, Zhiyuan Liu, Maosong Sun

TLDR: This paper investigates the phenomenon of 'complicit facilitation' in large language models (LLMs), where models provide guidance that enables illicit user actions. The authors develop a benchmark with 269 illicit scenarios and 50 intents based on real legal cases and legal frameworks to evaluate LLMs across socio-legal contexts. Findings show that GPT-4o provides illicit assistance in nearly half of the cases, and models often fail to deliver effective legal warnings or positive guidance. The study reveals significant variations in safety performance: higher complicity for crimes against societal interests, common non-extreme violations, and malicious intents with deceptive justifications. Socially, marginalized groups—including older adults, racial minorities, and those in lower-prestige jobs—are disproportionately targeted with unlawful responses. The paper links these biases to model-perceived stereotypes related to warmth and competence. Finally, it argues that current safety alignment techniques are inadequate and may worsen complicity.

PDF

18. Memories Retrieved from Many Paths: A Multi-Prefix Framework for Robust Detection of Training Data Leakage in Large Language Models

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20799

Authors: Trung Cuong Dang, David Mohaisen

TLDR: This paper introduces a multi-prefix framework for detecting training data leakage in large language models (LLMs). The authors argue that memorized sequences are deeply encoded and can be retrieved through a wide variety of distinct prefixes, unlike non-memorized content. They define memorization based on the ability of an adversarial search to identify a target number of distinct prefixes that trigger the same sequence, thus measuring the robustness of memory via retrieval path diversity. The method is tested on both open-source and aligned chat models, showing improved reliability in distinguishing memorized from non-memorized data, offering a practical tool for auditing privacy and copyright risks in LLMs.

PDF

19. Structured Prompting Enables More Robust, Holistic Evaluation of Language Models

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20836

Authors: Asad Aali, Muhammad Ahmed Mohsin, Vasiliki Bikia, Arnav Singhvi, Richard Gaus, Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Yifan Mai, Jordan Cahoon, Michael Pfeffer, Roxana Daneshjou, Sanmi Koyejo, Emily Alsentzer, Percy Liang, Christopher Potts, Nigam H. Shah, Akshay S. Chaudhari

TLDR: This paper introduces a reproducible DSPy+HELM framework that integrates structured prompting methods—such as chain-of-thought—into large-scale language model (LM) benchmarking. It addresses the limitations of traditional benchmarking frameworks like HELM, which rely on fixed prompts and often underestimate LM performance due to poor generalization across models. By using scalable, optimized structured prompts via DSPy, the study demonstrates that performance estimates become more robust, consistent, and representative across diverse benchmarks (including general and medical domains). Key findings include: (i) HELM underestimates LM performance by an average of 4%, (ii) performance variance across benchmarks increases without structured prompting, (iii) leaderboard rankings change on 3 out of 7 benchmarks, and (iv) reasoning-based prompting reduces sensitivity to prompt design. The authors open-source their integration and optimization pipeline, enabling reproducible and more accurate LM evaluation. This work highlights the importance of prompt optimization in achieving reliable performance ceilings for decision-making in LM deployment.

PDF

20. Winning with Less for Low Resource Languages: Advantage of Cross-Lingual English_Persian Argument Mining Model over LLM Augmentation

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20872

Authors: Ali Jahan, Masood Ghayoomi, Annette Hautli-Janisz

TLDR: This paper investigates cross-lingual argument mining for low-resource languages, focusing on Persian, by comparing three training strategies: zero-shot transfer from English, English-only training enhanced with LLM-generated synthetic data, and a cross-lingual model trained on both English and manually translated Persian data. Experiments on the English Microtext corpus and its Persian translation show that the cross-lingual model achieves the highest F1 score (74.8%) on Persian, outperforming both zero-shot transfer and LLM augmentation. The results suggest that combining native data from multiple languages, even with minimal translation effort, is more effective than relying on LLM-generated synthetic data for low-resource language tasks, offering a practical and efficient solution for argument mining in data-scarce settings.

PDF

21. A Unified Understanding of Offline Data Selection and Online Self-refining Generation for Post-training LLMs

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21056

Authors: Quan Xiao, Tianyi Chen

TLDR: This paper presents a unified framework for offline data selection and online self-refining generation in post-training large language models (LLMs). It formulates offline data selection as a bilevel optimization problem to improve data quality based on validation performance, while treating online self-refining generation as a model adaptation step that selects the best-performing model responses. The approach assigns learned weights to questions and responses, either explicitly or implicitly, to enhance the effectiveness of fine-tuning. The authors provide theoretical justification for the bilevel data selection method and demonstrate its superiority over unfiltered baselines. Experiments show improved performance in both quality enhancement and safety-aware fine-tuning, highlighting the benefits of combining validation-weighted offline data with online self-refining generations.

PDF

22. Can Finetuing LLMs on Small Human Samples Increase Heterogeneity, Alignment, and Belief-Action Coherence?

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21218

Authors: Steven Wang, Kyle Hunt, Shaojie Tang, Kenneth Joseph

TLDR: This paper investigates whether fine-tuning large language models (LLMs) on small human survey datasets—such as those from pilot studies—can improve the realism of LLM-generated responses in behavioral research. The study focuses on three key aspects: heterogeneity, alignment with human behavior, and belief-action coherence. Using an information disclosure experiment, the authors compare human responses with those generated by both base and fine-tuned LLMs. Results show that fine-tuning significantly enhances heterogeneity, alignment, and belief-action coherence compared to the base model. However, even the best fine-tuned models fail to replicate the regression coefficients observed in real human data, indicating that LLM-generated data are still unsuitable for formal inferential statistical analysis. The findings suggest that while fine-tuning on small human samples improves some aspects of simulation realism, it does not fully bridge the gap between synthetic and human behavior.

PDF

23. PEFT-Bench: A Parameter-Efficient Fine-Tuning Methods

Benchmark

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21285

Authors: Robert Belanec, Branislav Pecher, Ivan Srba, Maria Bielikova

TLDR: This paper introduces PEFT-Bench, a comprehensive and reproducible benchmark for evaluating parameter-efficient fine-tuning (PEFT) methods in autoregressive large language models (LLMs). It addresses the limitations of existing evaluations by covering 27 NLP datasets and 6 PEFT methods, and proposes a novel metric called PEFT Soft Score Penalties (PSCP) that incorporates trainable parameters, inference speed, and training memory usage. The benchmark aims to facilitate fair and holistic comparisons of PEFT techniques, supporting more sustainable and accessible LLM adaptation.

PDF

24. TALES: A Taxonomy and Analysis of Cultural Representations in LLM-generated Stories

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21322

Authors: Kirti Bhagat, Shaily Bhatt, Athul Velagapudi, Aditya Vashistha, Shachi Dave, Danish Pruthi

TLDR: This paper introduces TALES, a framework for evaluating cultural misrepresentations in stories generated by large language models (LLMs), with a focus on diverse Indian cultural identities. The authors develop TALES-Tax, a taxonomy of cultural inaccuracies derived from insights gathered through focus groups (N=9) and surveys (N=15) involving individuals with lived experiences in India. Using this taxonomy, they conduct a large-scale annotation study involving 108 annotators from 71 regions and 14 languages across India, analyzing 2,925 annotations from 6 LLMs. The results reveal that 88% of generated stories contain at least one cultural inaccuracy, with higher error rates in mid- and low-resourced languages and stories set in peri-urban areas. Surprisingly, despite these inaccuracies, models often possess the underlying cultural knowledge, as demonstrated by the creation of TALES-QA—a question bank to assess cultural knowledge—which shows strong performance in factual recall. The study highlights critical gaps in culturally accurate generation and offers tools for evaluating and improving cultural representation in LLMs.

PDF

25. RoParQ: Paraphrase-Aware Alignment of Large Language Models Towards Robustness to Paraphrased Questions

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21568

Authors: Minjoon Choi

TLDR: This paper introduces RoParQ, a benchmark designed to evaluate the consistency of Large Language Models (LLMs) in answering paraphrased questions in closed-book multiple-choice question answering tasks. The benchmark is created by generating paraphrases using proprietary models and selecting instances where a judge model shows inconsistent confidence. The authors propose XParaCon, a metric that measures robustness by calculating the standard deviation of accuracy across paraphrased variants. To improve consistency, they develop a reasoning-based, paraphrase-aware Supervised Fine-Tuning (SFT) strategy that promotes semantic invariance. Experimental results show that this method significantly improves model robustness, enabling lightweight models to achieve consistency levels comparable to larger pre-trained models, thus reducing reliance on surface-level patterns.

PDF

26. Beyond URLs: Metadata Diversity and Position for Efficient LLM Pretraining

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21613

Authors: Dongyang Fan, Diba Hashemi, Sai Praneeth Karimireddy, Martin Jaggi

TLDR: This paper explores the use of diverse metadata types—beyond just URLs—for improving the efficiency of Large Language Model (LLM) pretraining. The authors find that metadata with fine-grained information about document quality, when prepended to text, can accelerate training. They introduce 'metadata appending' as a strategy where predicting metadata serves as an auxiliary task, enhancing training speed. Learnable meta-tokens trained with masked language modeling also recover part of the efficiency gain by fostering quality-aware latent representations. Through probing analysis, the study reveals how metadata influences model learning. The findings provide practical guidance for integrating metadata to boost both training efficiency and model effectiveness in LLM pretraining.

PDF

27. Revisiting Generalization Across Difficulty Levels: It's Not So Easy

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21692

Authors: Yeganeh Kordi, Nihal V. Nayak, Max Zuo, Ilana Nguyen, Stephen H. Bach

TLDR: This paper investigates how well large language models (LLMs) generalize across different levels of task difficulty, focusing on the implications for data curation and evaluation. The authors use a large-scale, objective approach to rank example difficulty across six datasets by leveraging the outputs of thousands of LLMs and Item Response Theory (IRT), a method widely used in educational assessment. By avoiding human judgments of difficulty, the study provides a more consistent and scalable difficulty metric. The findings reveal that generalization across difficulty levels is often limited: training on either easy or hard data does not consistently improve performance across all difficulty levels. The paper emphasizes the need for diverse difficulty levels in both training and evaluation data, cautioning against simplifying data selection by focusing on only easy or hard examples.

PDF

28. Scaling Efficient LLMs

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2402.14746

Authors: B. N. Kausik

TLDR: This paper investigates efficient large language models (LLMs) by challenging the conventional 'AI scaling law' that suggests a linear relationship between model parameters and training data size. The authors derive a new scaling law where the number of parameters scales as D^{γ} with D being the data size and $ \gamma \in [0.44, 0.72]$, indicating potential for more parameter-efficient models. To realize this, they propose 'recurrent transformers'—a novel architecture that applies a single transformer layer iteratively over a sliding window of fixed width. This design enables linear time complexity in sequence length, memory efficiency, support for both forgetting and long-range memory (e.g., copy tasks), and compatibility with curriculum learning to mitigate vanishing gradients. Experimental results show strong performance on benchmarks, demonstrating the viability of this efficient approach.

PDF

29. Beyond Introspection: Reinforcing Thinking via Externalist Behavioral Feedback

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2501.01457

Authors: Diji Yang, Linda Zeng, Kezhen Chen, Yi Zhang

TLDR: This paper introduces the Distillation-Reinforcement-Reasoning (DRR) framework, an externalist approach to improving the reliability of reasoning in Large Language Models (LLMs). Instead of relying on self-critique, which suffers from the introspection illusion, DRR uses an external Discriminative Model (DM) trained on behavioral traces of the LLM's reasoning process. The DM evaluates observable reasoning steps at inference time, identifying and rejecting flawed pathways, thereby guiding the LLM toward better reasoning without modifying the base model. The method is lightweight, annotation-free, and effective across multiple reasoning benchmarks, outperforming existing self-critique techniques. The approach draws inspiration from ethological methods, emphasizing observable behavior over internal introspection.

PDF

30. BoundingDocs: a Unified Dataset for Document Question Answering with Spatial Annotations

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2501.03403

Authors: Simone Giovannini, Fabio Coppini, Andrea Gemelli, Simone Marinai

TLDR: This paper introduces BoundingDocs, a unified dataset for document question answering (QA) that integrates multiple public datasets related to Document AI and visually rich document understanding (VRDU). The key contributions are: (1) reformulating traditional Document AI tasks like Information Extraction (IE) into QA format, making them suitable for training and evaluating large language models (LLMs); and (2) providing OCR text and precise spatial annotations (bounding boxes) for answer locations in document images. The authors use this dataset to study how different prompting strategies—especially those incorporating spatial information—impact the performance of open-weight LLMs in document comprehension, identifying effective techniques for leveraging visual and spatial context.

PDF

31. Web-Shepherd: Advancing PRMs for Reinforcing Web Agents

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2505.15277

Authors: Hyungjoo Chae, Sunghwan Kim, Junhee Cho, Seungone Kim, Seungjun Moon, Gyeom Hwangbo, Dongha Lim, Minjin Kim, Yeonjun Hwang, Minju Gwak, Dongwook Choi, Minseok Kang, Gwanhoon Im, ByeongUng Cho, Hyojun Kim, Jun Hee Han, Taeyoon Kwon, Minju Kim, Beong-woo Kwak, Dongjin Kang, Jinyoung Yeo

TLDR: This paper introduces Web-Shepherd, the first Process Reward Model (PRM) designed specifically for web navigation tasks, enabling step-level evaluation of web agent trajectories. To support this, the authors create WebPRM Collection, a large-scale dataset with 40,000 step-level preference pairs and annotated checklists across diverse domains and difficulty levels. They also propose WebRewardBench, the first meta-evaluation benchmark for PRMs. Experiments show Web-Shepherd outperforms GPT-4o by about 30 points on the benchmark and improves performance by 10.9 points over using GPT-4o-mini as a verifier in WebArena-lite, while reducing cost by 10 times. The model, dataset, and code are publicly available.

PDF

32. UITron-Speech: Towards Automated GUI Agents Based on Speech Instructions

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2506.11127

Authors: Wenkang Han, Zhixiong Zeng, Jing Huang, Shu Jiang, Liming Zheng, Longrong Yang, Haibo Qiu, Chang Yao, Jingyuan Chen, Lin Ma

TLDR: This paper introduces UITron-Speech, an end-to-end GUI agent that processes speech instructions and on-device screenshots to perform user actions, aiming to enhance accessibility in hands-free scenarios. To overcome data scarcity, the authors synthesize high-quality speech instruction datasets using a random-speaker text-to-speech model. A mixed-modality training strategy addresses modality imbalance in foundation models, and a training-free two-step grounding refinement method reduces localization errors. Experiments show strong performance and adaptability across benchmarks, demonstrating the feasibility of speech-driven GUI agents for more inclusive human-computer interaction. Code and datasets are publicly available.

PDF

33. AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2507.00310

Authors: Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, Peter Clark

TLDR: This paper introduces AutoDiscovery, a method for open-ended scientific discovery (ASD) that uses Bayesian surprise to guide autonomous hypothesis generation and exploration. Unlike traditional approaches that rely on human-specified research questions, AutoDiscovery enables AI to identify novel and surprising hypotheses by measuring the epistemic shift in a large language model's (LLM) beliefs after observing experimental results. The system employs Monte Carlo Tree Search (MCTS) with progressive widening, using surprisal as the reward function to efficiently navigate complex hypothesis spaces. Evaluated on 21 real-world datasets across biology, economics, finance, and behavioral science, AutoDiscovery achieves 5–29% more surprising discoveries than baseline methods under a fixed budget. Human evaluations confirm that two-thirds of the system's discoveries are also surprising to domain experts, indicating strong potential for advancing autonomous scientific exploration. The work highlights a promising direction for AI-driven discovery that aligns with the goals of open-ended, self-directed research.

PDF

34. On The Role of Pretrained Language Models in General-Purpose Text Embeddings: A Survey

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2507.20783

Authors: Meishan Zhang, Xin Zhang, Xinping Zhao, Shouzheng Huang, Baotian Hu, Min Zhang

TLDR: This survey paper explores the role of pretrained language models (PLMs) in the development of general-purpose text embeddings (GPTE). It examines how PLMs contribute to GPTE through embedding extraction, enhancing representational expressivity, shaping training strategies, defining learning objectives, and constructing training data. The paper also highlights advanced capabilities enabled by PLMs, such as multilingual support, multimodal integration, code understanding, and adaptation to specific scenarios. Finally, it outlines future research directions, including integrating ranking mechanisms, improving safety, mitigating bias, incorporating structural information, and extending embeddings to support cognitive tasks. The paper provides a comprehensive overview of GPTE's evolution and potential in the era of PLMs.

PDF

35. Where to Start Alignment? Diffusion Large Language Model May Demand a Distinct Position

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2508.12398

Authors: Zhixin Xie, Xurui Song, Jun Luo

TLDR: This paper presents the first safety analysis of Diffusion Large Language Models (dLLMs), a non-autoregressive paradigm that generates text through diffusion processes. The authors identify a critical security asymmetry: while middle tokens in the output are more crucial for safety, attackers have limited ability to manipulate them due to the model's inherent sequential generation tendency. To address this, the authors propose Middle-tOken Safety Alignment (MOSA), a reinforcement learning-based method that directly aligns middle tokens with safe refusals. Experiments show MOSA significantly improves safety against eight attack methods across two benchmarks, while maintaining strong performance in coding, math, and general reasoning tasks.

PDF

36. Uncovering Implicit Bias in Large Language Models with Concept Learning Dataset

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2510.01219

Authors: Leroy Z. Wang

TLDR: This paper introduces a concept learning dataset designed to reveal implicit biases in large language models (LLMs). Through in-context concept learning experiments, the authors identify a bias toward upward monotonicity in quantifiers—meaning models tend to interpret quantified statements in a way that favors broader or more inclusive interpretations. This bias is more evident in concept learning scenarios than in direct prompting, suggesting that in-context learning can expose hidden biases that are not apparent in standard testing. The study highlights the importance of using concept learning frameworks to evaluate and understand model behavior beyond surface-level performance.

PDF

37. BengaliFig: A Low-Resource Challenge for Figurative and Culturally Grounded Reasoning in Bengali

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20399

Authors: Abdullah Al Sefat

TLDR: This paper introduces BengaliFig, a low-resource dataset designed to evaluate figurative and culturally grounded reasoning in Bengali, a widely spoken but under-resourced language. The dataset consists of 435 riddles from Bengali oral and literary traditions, annotated across five dimensions: reasoning type, trap type, cultural depth, answer category, and difficulty. It is converted into a multiple-choice format using an AI-assisted, constraint-aware pipeline. The study evaluates eight state-of-the-art LLMs using zero-shot and few-shot chain-of-thought prompting, revealing persistent weaknesses in metaphorical and culturally specific reasoning. BengaliFig serves as a diagnostic tool for assessing LLM robustness in culturally rich, low-resource settings and promotes more inclusive NLP evaluation.

PDF

38. Structured Definitions and Segmentations for Legal Reasoning in LLMs: A Study on Indian Legal Data

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20669

Authors: Mann Khatri, Mirza Yusuf, Rajiv Ratn Shah, Ponnurangam Kumaraguru

TLDR: This paper investigates the challenges of applying Large Language Models (LLMs) to legal reasoning, particularly in the context of Indian legal judgments. It explores how structured data organization, definition of rhetorical roles, and step-by-step reasoning emulation can improve LLM performance in legal tasks. The study conducts zero-shot experiments on three Indian legal judgment prediction datasets, demonstrating that reorganizing legal documents by rhetorical structure and defining key legal terms significantly enhances model accuracy, with F1 score improvements ranging from 1.5% to 4.36%. The findings highlight the importance of domain-specific data structuring and terminology clarification for improving LLMs in specialized domains like law.

PDF

39. Prompt Engineering Techniques for Context-dependent Text-to-SQL in Arabic

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20677

Authors: Saleh Almohaimeed, May Alsofyani, Saad Almohaimeed, Mansour Al Ghanim, Liqiang Wang

TLDR: This paper introduces Ar-SParC, the first Arabic cross-domain, context-dependent text-to-SQL dataset, containing 3,450 multi-turn question sequences (10,225 questions) with corresponding SQL queries. It evaluates 40 experiments using GPT-3.5-turbo and GPT-4.5-turbo with 10 prompt engineering techniques, including question representation and in-context learning methods. The authors propose a novel GAT corrector that improves execution and interaction accuracy across zero-shot and in-context learning settings, outperforming prior GAT verifier methods, especially for Arabic. An ablation study explains the effectiveness of the GAT corrector in the Arabic context.

PDF

40. Dynamic Template Selection for Output Token Generation Optimization: MLP-Based and Transformer Approaches

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20683

Authors: Bharadwaj Yadavalli

TLDR: This paper introduces Dynamic Template Selection (DTS), a method to optimize output token generation in large language models by adaptively choosing response templates based on query complexity. The approach aims to reduce token costs—especially for expensive output tokens—without sacrificing response quality. Two routing strategies are evaluated: a lightweight MLP using pre-computed embeddings and a more complex fine-tuned RoBERTa transformer. The MLP achieves 90.5% routing accuracy with significantly fewer parameters than RoBERTa (89.5% accuracy), demonstrating strong efficiency. The method shows strong generalization across three major LLM providers (OpenAI GPT-4, Google Gemini, Anthropic Claude), achieving 32.6%–33.9% token reductions. The work includes a formal problem formulation, algorithmic design with complexity analysis, and extensive empirical validation in production settings.

PDF

41. ST-PPO: Stabilized Off-Policy Proximal Policy Optimization for Multi-Turn Agents Training

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20718

Authors: Chenliang Li, Adel Elmahdy, Alex Boyd, Zhongruo Wang, Alfredo Garcia, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, Mingyi Hong

TLDR: This paper introduces ST-PPO, a stabilized off-policy Proximal Policy Optimization method designed for training multi-turn language model agents. The authors identify two key sources of instability in standard token-level PPO: misaligned token-level importance sampling and inaccurate advantage estimates from off-policy samples. To address these, they propose two complementary techniques: turn-level importance sampling, which aligns optimization with the natural turn-based structure of multi-turn tasks, and clipping-bias correction, which reduces gradient variance by downweighting unreliable off-policy samples. The resulting ST-PPO combines both techniques and demonstrates superior stability and performance over standard PPO in multi-turn reasoning tasks across QA, multi-hop QA, and medical multiple-choice QA benchmarks. The method maintains lower clipping ratios and avoids performance collapse, offering a scalable solution for training LLM agents in complex, multi-turn environments.

PDF

42. SAGE: An Agentic Explainer Framework for Interpreting SAE Features in Language Models

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20820

Authors: Jiaojiao Han, Wujiang Xu, Mingyu Jin, Mengnan Du

TLDR: This paper introduces SAGE (SAE AGentic Explainer), an agent-based framework designed to interpret features extracted by Sparse Autoencoders (SAEs) in large language models (LLMs). SAGE transforms the traditionally passive process of feature explanation into an active, iterative, and explanation-driven approach. It formulates multiple hypotheses for each feature, designs targeted experiments to test these hypotheses, and refines explanations using empirical activation feedback. The framework demonstrates superior performance in both generating and predicting feature behaviors across diverse language models, outperforming existing state-of-the-art methods in interpretability and accuracy.

PDF

43. Length-MAX Tokenizer for Language Models

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20849

Authors: Dong Dong, Weijie Su

TLDR: This paper introduces the Length-MAX tokenizer, a novel approach to tokenization in language models that minimizes the average number of tokens per character by formulating the vocabulary construction as a length-weighted graph partitioning problem and solving it with a greedy algorithm. Compared to traditional Byte Pair Encoding (BPE), Length-MAX achieves 14–18% fewer tokens across various vocabulary sizes (10K–64K), leading to reduced training steps (17–18.5%), lower inference latency (12.7–13.7%), and 16% higher throughput at 124M parameters. It also improves downstream performance (e.g., 11.7% lower LAMBADA perplexity, 4.3% higher HellaSwag accuracy) while maintaining high vocabulary coverage (99.62%) and low out-of-vocabulary rates (0.12%). The tokenizer is compatible with production systems and reduces embedding and KV-cache memory by 18% at inference, making it highly efficient for both training and deployment.

PDF

44. Evo-Memory: Benchmarking LLM Agent Test-time Learning with Self-Evolving Memory

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20857

Authors: Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, Derek Zhiyuan Cheng

TLDR: This paper introduces Evo-Memory, a benchmark and framework for evaluating self-evolving memory in large language model (LLM) agents during test-time learning. It addresses the limitation of existing evaluations that focus on static, conversational settings by proposing a streaming benchmark with sequential task streams that require LLMs to continuously update and reuse memory. The framework integrates over ten memory modules and evaluates them across diverse datasets involving multi-turn goal-oriented tasks and reasoning/QA. The authors propose ExpRAG as a baseline for experience reuse and ReMem, a pipeline that tightly couples reasoning, actions, and memory refinement to enable continual improvement. The work emphasizes dynamic memory management in real-world applications like interactive problem-solving assistants.

PDF

45. Chatty-KG: A Multi-Agent AI System for On-Demand Conversational Question Answering over Knowledge Graphs

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.20940

Authors: Reham Omar, Abdelghny Orogat, Ibrahim Abdelaziz, Omij Mangukiya, Panos Kalnis, Essam Mansour

TLDR: Chatty-KG proposes a modular multi-agent system for conversational question answering over knowledge graphs (KGs), combining the strengths of large language models (LLMs) and structured KGs. It addresses the limitations of existing approaches—such as RAG systems that serialize graph structure and traditional KGQA systems that struggle with multi-turn dialogue—by using task-specialized LLM agents to interpret context, track dialogue, link entities and relations, and generate accurate SPARQL queries. The system enables low-latency, context-aware, and scalable multi-turn QA without requiring fine-tuning or pre-processing. Evaluations on diverse KGs show significant improvements over state-of-the-art methods in F1 and P@1 scores, with strong compatibility across commercial (e.g., GPT-4o, Gemini-2.0) and open-weight (e.g., Phi-4, Gemma 3) LLMs. The design supports evolving KGs and maintains dialogue coherence, making it suitable for dynamic, real-world applications.

PDF

46. TrackList: Tracing Back Query Linguistic Diversity for Head and Tail Knowledge in Open Large Language Models

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21006

Authors: Ioana Buhnila, Aman Sinha, Mathieu Constant

TLDR: This paper introduces TrackList, a fine-grained linguistic and statistical analysis framework to examine how pre-training data influences Large Language Models' (LLMs) performance across different types of queries, particularly focusing on definition, exemplification, and paraphrasing. Using the newly created RefoMed-EN dataset with 6,170 human-annotated medical terms, the study evaluates LLMs' responses to head (frequent) and tail (rare) knowledge. Results indicate that LLMs perform best on definition-type queries and significantly worse on exemplification tasks. Furthermore, models tend to paraphrase more frequently for common, high-frequency concepts, especially in expert-level texts, while underperforming on rare or technical terms. The findings highlight a bias in LLMs toward frequent knowledge and suggest limitations in handling diverse linguistic expressions, particularly for low-frequency or specialized content.

PDF

47. Semantic Anchors in In-Context Learning: Why Small LLMs Cannot Flip Their Labels

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21038

Authors: Anantha Padmanaban Krishna Kumar (Boston University)

TLDR: This paper investigates the fundamental limits of in-context learning (ICL) in large language models (LLMs) by examining whether ICL can override pre-trained label semantics or merely refine an existing semantic backbone. The authors treat LLMs as prompt-induced classifiers and compare model behavior under natural demonstrations (correct labels) versus inverted demonstrations (flipped label meanings). They introduce three alignment metrics—truth, prior, and prompt alignment—and define a semantic override rate to measure correctness under flipped semantics. Across eight classification tasks and eight open-source LLMs (1–12B parameters), the results consistently support a 'semantic anchor' view: ICL does not enable models to learn coherent anti-semantic classifiers. Even with inverted demonstrations, models fail to achieve semantic override, showing zero semantic override rates in few-shot settings. Instead, ICL primarily adjusts how inputs project onto stable semantic directions established during pre-training, suggesting that label semantics are deeply anchored and difficult to override through prompting alone. The findings imply that effective semantic remapping at scale may require interventions beyond standard ICL.

PDF

48. Orthographic Constraint Satisfaction and Human Difficulty Alignment in Large Language Models

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21086

Authors: Bryan E. Tuck, Rakesh M. Verma

TLDR: This paper investigates how large language models handle orthographic constraints—specifically, the ability to generate text that adheres to precise character-level rules, such as in word puzzles. The study evaluates 28 model configurations across three families (Qwen3, Claude Haiku-4.5, GPT-5-mini) on 58 puzzles requiring strict orthographic compliance. Results show significant performance differences between architectures (F1 scores ranging from 0.343 to 0.761), with architectural design having a greater impact than parameter scaling. High-capacity models benefit more from increased thinking budget, while mid-sized models show diminishing or negative returns. The models exhibit modest alignment with human difficulty (r=0.24–0.38), but consistently fail on common words with unusual spelling (e.g., 'data', 'poop', 'loll'), despite high human success rates. These failures suggest models prioritize distributional plausibility over orthographic correctness, indicating a need for architectural or training innovations beyond scaling.

PDF

49. How to Correctly Report LLM-as-a-Judge Evaluations

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21140

Authors: Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, Kangwook Lee

TLDR: This paper addresses the challenges of using large language models (LLMs) as evaluators in place of human judges, focusing on the noise and bias in LLM-based evaluations due to imperfect specificity and sensitivity. It proposes a plug-in framework that corrects bias and constructs confidence intervals that account for uncertainty from both test and calibration datasets. The method is practical and statistically sound, even when only estimated values of model performance are available. Additionally, the paper introduces an adaptive algorithm to optimally allocate calibration samples, reducing uncertainty in accuracy estimates. The work aims to improve the reliability and validity of LLM-as-a-judge evaluation practices.

PDF

50. AnchorOPT: Towards Optimizing Dynamic Anchors for Adaptive Prompt Learning

Relevance: ⭐⭐⭐⭐⭐

arXiv ID: 2511.21188

Authors: Zheng Li, Yibing Song, Xin Zhang, Lei Luo, Xiang Li, Jian Yang

TLDR: This paper proposes AnchorOPT, a dynamic anchor-based prompt learning framework designed to enhance the adaptability of prompt learning in vision-language models like CLIP. Unlike traditional methods that use static textual anchors (e.g., 'shape', 'color'), AnchorOPT learns anchor values dynamically from task-specific data and introduces a learnable position matrix that adaptively adjusts the positional relationship between anchors and soft tokens based on task context and training stage. The method operates in two stages: first learning anchors, then freezing them while optimizing soft tokens and the position matrix. Experiments show that AnchorOPT achieves performance on par with or better than methods using more complex modules, while being easily integrable as a plug-and-play component. The approach improves generalization and adaptability across tasks and datasets.

PDF