Large Language Models Introduction

BHARAT BHARGAVA

PURDUE UNIVERSITY

LLM

Modern Large Language Models(LLM):

- LLM is pretrained on massive corpus of text data
- Common Tasks Performed by LLMs: Content Generation, Information Retrieval and Summary, Translation & Rewriting, Chatbots and Conversational AI, Agents and Tool-Calling Tasks

Architecture:

- Encoder-Decoder, Decoder Only

Instruction Tuning and RLHF:

- Instruction Tuned models are finetuned on (instruction, response) pairs
- Reinforcement Learning on Human Feedback (RLHF) finetunes models on human feedback

Challenges of LLM

Hallucination:

- LLM tends to generate factually incorrect or unsupported answers

Explainability:

- Difficult to interpret decisions from LLM

Bias and Fairness:

- LLMs tends to reflect and amplify societal or training-data biases

Resource Intensive:

- Pretraining of LLMs is data and computation intensive

Fundamentals

RNN:

- A class of neural network designed to handle sequential data by using internal memory (a hidden state) to process a sequence of inputs

LSTM:

- Recurrent neural network (RNN) that uses a gating mechanism to selectively retain or forget information over long sequences, solving the vanishing gradient problem

Attention:

- A mechanism that allows a neural network to dynamically weigh the importance of different parts of the input data when producing an output

Multilayer Perceptron (MLP):

- A foundational feedforward artificial neural network composed of at least three layers (input, output, and one or more hidden layers) where neurons are fully connected, and activation functions introduce non-linearity

Fundamentals

Embedding:

- Low-dimensional, dense, real-valued vector representation of discrete objects (like words, categories, or entities) that captures their semantic relationships and allows them to be efficiently processed by machine learning models.

Continual Pretraining:

- An adaptation method where a language model, already pretrained on a large general corpus, is further trained iteratively on a continuously updating or domain-specific dataset to maintain or improve its performance over time without forgetting previous knowledge.

Supervised Learning:

- paradigm where the model is trained on labeled data (input-output pairs) to learn a mapping function that can predict the output for new, unseen inputs.

Contextual Multi-Armed Bandit:

- An online learning problem where an agent chooses one of several actions (arms) based on a current state (context) to maximize cumulative reward, balancing exploration of unknown actions with exploitation of known good actions.

Fundamentals

Transformer:

- A neural network architecture that relies on a self-attention mechanism to process all input data simultaneously, making it highly effective for tasks like machine translation and language understanding.

Convolutional Neural Network (CNN):

- A type of deep neural network that uses learnable filters (or kernels) to automatically detect and learn hierarchical patterns in grid-like data, such as images.

Generative Adversarial Network (GAN):

- A framework consisting of two competing neural networks, a generator and a discriminator, which are trained together to create realistic synthetic data that mimics a given distribution.

Autoencoder:

- An unsupervised neural network that learns a compressed representation (encoding) of input data and then reconstructs the original data from this encoding, used for feature learning and dimensionality reduction.

Fundamentals

Foundation Model:

- A large-scale model trained on a vast quantity of broad data that can be adapted to a wide range of downstream tasks, serving as a base for more specialized applications

Meta-Learning:

- An advanced subfield of machine learning, often called "learning to learn," where the goal is to train a model on a variety of tasks so it can solve new, unseen tasks more efficiently

Multi-arm Bandit Fundamentals

Upper Confidence Bound (UCB):

- Achieves a balance by selecting the arm that maximizes the potential for high reward, not just the observed average reward

Adaptive Greedy (AG):

- Explores by choosing the action with the highest potential reward, factoring in both its estimated performance and the uncertainty around that estimate

Epsilon Greedy (EG):

- chooses the best-known option, but with a small, fixed probability (ϵ), it deliberately chooses a completely random action to discover other possibilities

Reward

- a scalar feedback signal given by the environment to the agent after each action, indicating the immediate desirability of that action and serving as the primary driver for learning the optimal behavior policy

Can AI Expand Human Minds

Classic dual-process model of human decision-making (Proposed by Daniel Kahneman, Nobel Prize in Economics)

- System 1 (Intuitive)
- System 2 (Analytical):

System 0 :

- Interacts with and augments both intuitive (System 1) and analytical (System 2)
- Fundamentally changes the starting point for human decision-making by curating the available information

Language Models Wrestle with Gaps in Understanding

The Core Challenge of Reliability

- LLMs exhibit "jagged intelligence," working well one moment and failing unexpectedly the next
- Even simple changes to a query can cause the AI to provide wildly incorrect answers.

Building Internal "World Models":

- Evidence suggests LLMs do more than statistical matching, creating abstract "world models."

Peeking Inside the Black Box:

- Researchers use machine learning probes to analyze an LLM's internal "state of mind"
- These probes show information is processed in logical "hops" through the model's layers

Strategies for Making LLMs More Reliable

Improving the Prompting Process

- Chain-of-Thought (CoT) prompting decomposes complex problems into a sequence of simpler steps
- New systems can now automate CoT prompting and backtrack from failures to find alternative paths

Combining LLMs with Formal Tools:

- The "LLM-Modulo" architecture combines language models with formal, symbolic tools for verification

The Challenge of Truth and Uncertainty:

- Researchers use machine learning probes to analyze an LLM's internal "state of mind"

Not Every AI Problem Is a Data Problem

Challenging the Scaling Narrative:

- Scaling isn't a universal solution; simply "throwing" more data and compute at every problem is not always effective

The Data Quality Bottleneck:

- We are running out of good data; as models grow, the finite supply of high-quality, human-generated data becomes a major constraint.
- Low-quality data is harmful; larger models are very sensitive to bad data and can memorize outliers, leading to undesirable outputs like the "glue on pizza" response
- Identifying data quality is necessary; models need some low-quality data to learn how to distinguish it from high-quality information and fix mistakes

Is Synthetic Data the Solution?:

- Synthetic data works in some areas; it has driven improvements in fields with automatic verification, such as math and coding

Intentional Scaling

Where Data-Driven Scaling Thrives:

- Success stories exist in stable domains; scaling has worked well for machine translation, robotics, and drug discovery
- Stable fields benefit from vast amounts of high-quality data with consistent, abstractable rules and patterns

Where Scaling Stumbles:

- LLM failures in robust reasoning are likely due to architectural issues, not a lack of data

Iterative Self-Checking LLMs with Constraint Programming for Robust TSP Optimization

PURDUE UNIVERSITY

Background

The Problem: LLMs Struggle with Complex, Constrained Problems:

- The Travelling Salesman Problem (TSP) is a foundational optimization challenge with wide real-world applications in logistics and manufacturing.
- While Large Language Models (LLMs) show promise in solving combinatorial tasks, they have significant limitations
  - Black-Box Nature: LLM reasoning is often opaque, making it difficult to trust or verify their solutions.
  - Constraint Violations: They can produce solutions that seem plausible but violate fundamental rules, such as skipping a city or not returning to the start.
  - Lack of Interpretability: It is hard to understand the reasoning behind a given solution, which is critical for practical deployment.

A Hybrid LLM + CP Framework

Constraint Programming (CP):

- Rigor of Constraint Programming (CP) can augment the pure LLM approach

LLM In-the-Loop:

- Use the LLM to generate creative strategies and candidate solutions, while the CP module acts as a strict enforcer of all problem rules and constraints

Self Correcting Approach:

- Self-correcting system that ensures solutions are not only feasible but are progressively improved through automated self-checking

Proposed Methodology

Continual Pretraining for Enhanced Reasoning:

- Datasets include: Mathematical Problem-Solving: GSM8K, MATH, and MR-BEN-math to improve multi-step computation and algebraic manipulation.
- General Reasoning & Logic: LogiQA, ProofWriter, and WinoGrande to enhance systematic thinking and deductive reasoning.

LLM-Powered Solution Generation

- To guarantee correctness, the LLM's outputs are validated by a CP module.

Integration with Constraint Programming (CP)

- To guarantee correctness, the LLM's outputs are validated by a CP module.

Iterative Self-Checking and Refinement Loop

- If an error or a suboptimal step is found, the LLM revises its proposal, and the process repeats until a feasible, high-quality solution is found

Reference

- Chen, J., Chi, L., Peng, B., & Yuan, Z. (2024). HLLM: Enhancing Sequential Recommendations via Hierarchical Large Language Models for Item and User Modeling. arXiv:2409.12740v1 [cs.IR]
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers fromaLanguageModel.InAdvancesinNeuralInformationProcessingSystems(NeurIPS). https://github.com/likenneth/honest_llama
- Gagandeep Kaur and Amit Sharma. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of Big Data*, 10(1), 2023. doi: 10.1186/s40537-022-00680-6.
- Huijian Han, Zhiming Li, and Zongwei Li. Using machine learning methods to predict consumer confidence from search engine data. *Sustainability*, 15(4), 2023. ISSN 2071-1050. doi: 10.3390/su15043100. URL https://www.mdpi.com/2071-1050/15/4/3100.
- Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. Clues be- fore answers: Generation-enhanced multiple-choice QA. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3272–3287, Seattle, United States, July 2022. Asso- ciation for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.239. URL https://aclanthology.org/2022.naacl-main.239.
- Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. *ArXiv*, abs/2210.12353, 2022. URL https://api.semanticscholar.org/CorpusID:253098700.
- Akshay Chaturvedi, Onkar Pandit, and Utpal Garain. CNN for text-based multiple choice question answering. In *Proceedings of the 56th Annual Meeting of the Association for Com- putational Linguistics (Volume 2: Short Papers)*, pp. 272–277, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2044. URL https://aclanthology.org/P18-2044.
- Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. Mmm: Multi- stage multi-task learning for multi-choice reading comprehension, 2019.
- Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, and Guoping Hu. Convolutional spatial attention model for reading comprehension with multiple-choice questions. *Pro- ceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6276–6283, jul 2019. doi: 10.1609/aaai.v33i01.33016276. URL https://doi.org/10.1609%2Faaai.v33i01.33016276.
- Zixian Huang, Ao Wu, Yulin Shen, Gong Cheng, and Yuzhong Qu. When retriever-reader meets scenario-based multiple-choice questions. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID: 237364132.
- Doe, J. 2025. The Future of Computing. *Communications of the ACM* 69, 10 (Oct. 2025), 123–456. https://doi.org/10.1145/xxxxxxx.xxxxxxx.