PDDL PLANNING WITH PRE-TRAINED LARGE LANGUAGE MODELS

GPT-made summary

This paper explores using large language models (LLMs), like OpenAI's Codex, to address PDDL-based AI planning problems. The idea is to leverage LLMs' reasoning capabilities through few-shot prompting to solve planning tasks or guide traditional heuristic-based planners. Results show that while LLMs struggle to independently solve complex problems, they perform well in simpler and novel domains. Moreover, their outputs significantly improve the performance of heuristic planners, demonstrating the potential of hybrid approaches that combine LLMs with traditional planning techniques.

Attention Is All You Need

GPT-made summary

This paper introduces the Transformer model, a novel architecture for sequence transduction that relies entirely on attention mechanisms, eliminating recurrence and convolution. The idea is to use self-attention to capture global dependencies in sequences, enabling greater parallelization and efficiency. Results show state-of-the-art performance in machine translation tasks, with faster training times compared to previous models, and strong generalization to other tasks like English constituency parsing.

LLMs Can Understand Encrypted Prompt: Towards Privacy-Computing Friendly Transformers

GPT-made summary

This paper proposes a framework to enable private inference for large language models (LLMs) using homomorphic encryption (HE) and secure multi-party computation (MPC). By replacing computationally expensive operators in transformers (e.g., GELU, Softmax, LayerNorm) with privacy-friendly approximations, the approach achieves 5× faster computation and 80% lower communication costs compared to existing methods, with minimal accuracy loss. This demonstrates the feasibility of efficient, privacy-preserving LLM inference.

CAN LLMS EXPRESS THEIR UNCERTAINTY? AN EMPIRICAL EVALUATION OF CONFIDENCE ELICITATION IN LLMS

This paper evaluates methods for enabling large language models (LLMs) to express accurate confidence in their responses using black-box techniques. It introduces a framework combining human-inspired prompting, sampling, and aggregation strategies to elicit and improve verbalized confidence. Results reveal that while LLMs often display overconfidence, certain strategies, like consistency-based aggregation and multi-step reasoning prompts, improve calibration and failure prediction. However, black-box methods remain less effective compared to white-box approaches, especially on tasks requiring specialized knowledge, emphasizing the need for further research.

Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners

> 2307.01928v2.pdf

This paper introduces KNOWNO, a framework for uncertainty alignment in LLM-based robotic planning, leveraging conformal prediction to ensure statistical task completion guarantees while minimizing human help. The idea is to calibrate LLM confidence to detect ambiguity and trigger human intervention only when needed. Results across simulated and hardware setups show KNOWNO improves efficiency and autonomy, reducing human assistance by 10–24% compared to baselines, while consistently achieving user-specified task success levels.

Look Before You Leap: An Exploratory Study of Uncertainty Analysis for Large Language Models

> 2307.10236v4.pdf

This paper explores uncertainty estimation techniques to assess the reliability of large language models (LLMs) in both NLP and code generation tasks. By evaluating 12 methods across 17 LLMs, it finds sample-based methods are most effective for risk prediction, while perturbation-based methods show task-specific performance. The study highlights challenges in uncertainty estimation for high-performing models and proposes directions for improved trustworthiness in real-world LLM applications.

DELIVERAI: A DISTRIBUTED PATH-SHARING NETWORK BASED SOLUTION FOR THE LAST MILE FOOD DELIVERY PROBLEM

> 2311.02017v2.pdf

The paper "DeliverAI: A Distributed Path-Sharing Network Based Solution for the Last Mile Food Delivery Problem" proposes a reinforcement learning-based solution to address the challenges of last-mile food delivery.

It introduces DeliverAI, a system that dynamically optimizes routes through distributed path-sharing among delivery agents, aiming to minimize delivery times and reduce fleet congestion. The results show significant improvements in total travel distance and delivery delays, demonstrating enhanced efficiency, particularly in urban settings.

Benchmarking LLMs via Uncertainty Quantification

> 2401.12794v3.pdf

GPT-made summary

The study "Benchmarking LLMs via Uncertainty Quantification" explores how uncertainty quantification (UQ) can be applied to evaluate and compare large language models (LLMs). The authors propose novel metrics leveraging UQ to assess the reliability of LLMs, going beyond traditional accuracy benchmarks. Their findings reveal that models incorporating UQ produce more robust predictions, especially in low-confidence scenarios, offering a valuable framework for future benchmarking efforts.

EXPLORING AND BENCHMARKING PLANNING CAPABILITIES OF LARGE LANGUAGE MODELS

> 2406.13094v2.pdf

GPT-made summary

In "Exploring and Benchmarking Planning Capabilities of Large Language Models," the authors examine the ability of LLMs to solve task-planning problems across classical AI domains and natural language tasks. By using diverse benchmarks, the study evaluates the feasibility and optimality of plans generated by LLMs, highlighting that while these models show promise, they often need tailored fine-tuning and structured prompting to deliver consistent results.

An AI Planning Approach to Emergency Material Scheduling Using Numerical PDDL

> 125977151.pdf

GPT-made summary

The paper "An AI Planning Approach to Emergency Material Scheduling Using Numerical PDDL" presents an application of AI planning techniques for optimizing emergency material transportation. By employing numerical PDDL to model resource allocation and delivery route scheduling, the study achieves improved efficiency in emergency response times, successfully integrating numerical constraints with classical planning methods to enhance logistics operations.

The Needle In a Haystack Test

> needle_in_a_haystack_test.pdf

GPT-made summary

"The Needle in a Haystack Test" introduces a new benchmark designed to assess the reasoning abilities of LLMs by challenging them to identify relevant solutions amidst a sea of distractors. The test simulates real-world reasoning scenarios by embedding misleading data alongside correct information, revealing that while LLMs perform well in structured tasks, their reasoning capabilities decline significantly when faced with ambiguous or highly distractive contexts.