Addressing Hallucinations of LLMs

Active Detection and Mitigation
Under Review @ ICLR

Selective Prediction and Calibration

Selective Prediction in IID, OOD, Adv Settings
ACL '22

Post-Abstention
ACL '23

Selective Prediction Approach
ACL '22 @ Repl4NLP

Dealing with Novelties

NoveltyTask in Authorship Attribution
ACL '23

Open-World Monopoly
AAMAS '23

Open-World Monopoly
AAAI '22 Symposium

Tool to Measure Reliability

Reliability-Checklist Tool
Under Review @ AAAI Demo

Instruction Tuning

Super-Natural Instructions
EMNLP '22

Alignment with 'Difficult' Instances
Ongoing

LLM Efficient Decoding

Intermediate Layer Decoding
Preprint

Reasoning

Numerical Reasoning - NumGLUE
ACL '22

Feasibility Reasoning - FeasibilityQA
EACL '23

Logical Reasoning - LogicAttack
EMNLP '23

Logical Reasoning - LogicBench
Under Review @ ICLR

▪ ▪ ▪

Inference Efficiency

Model Cascading
EMNLP '22

Dynamic Reading for Open-domain QA
AAAI '23 @ KGNLP

Data Sample Efficiency

Unsupervised Natural Language Inference
ACL '22

Curriculum Learning in Multi-task Learning
NAACL '22 @ DeepLO

Evaluation Efficiency

Instance-Level Difficulty Analysis
ACL '22

Knowledge Indexing Efficiency

Information Pointers as Parametric Knowledge
Ongoing

Evaluation of LLMs

Questions without Definitive Answers
ACL '23 @ TrustNLP