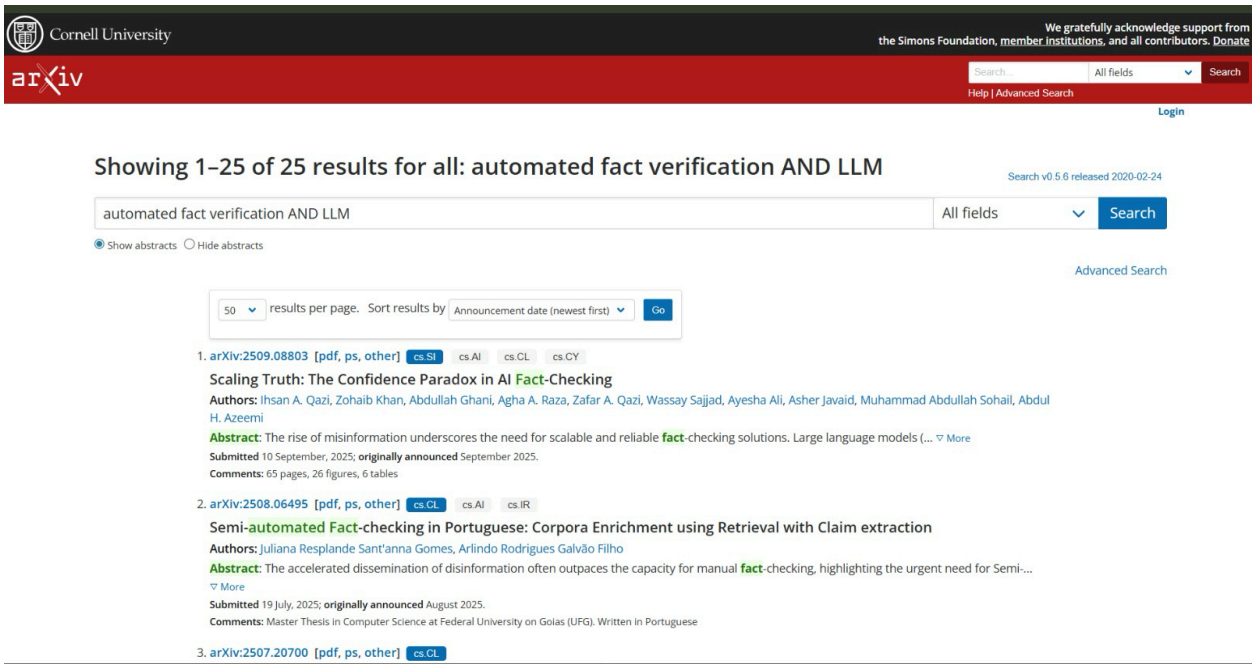
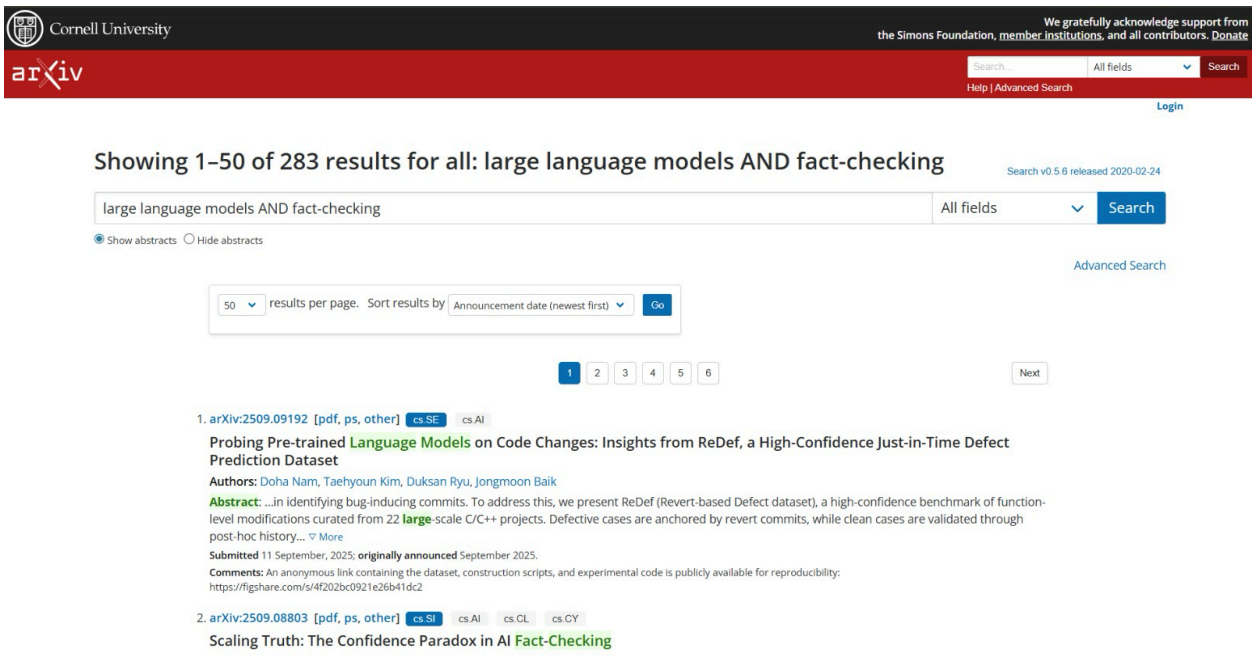
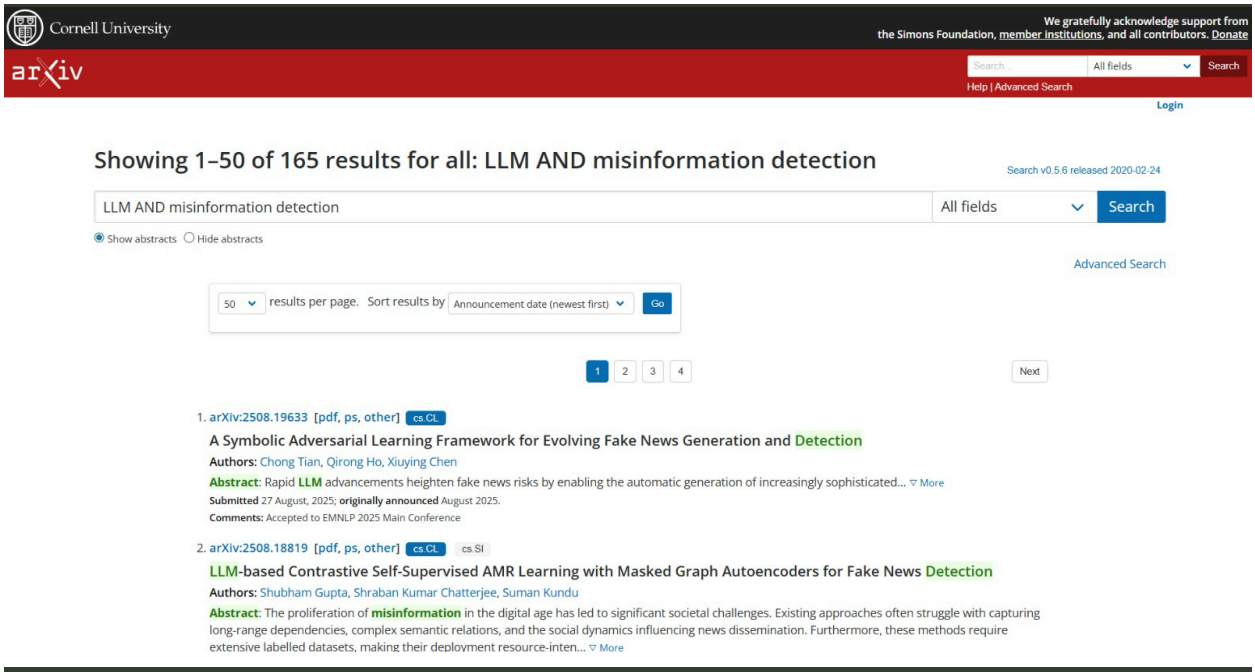
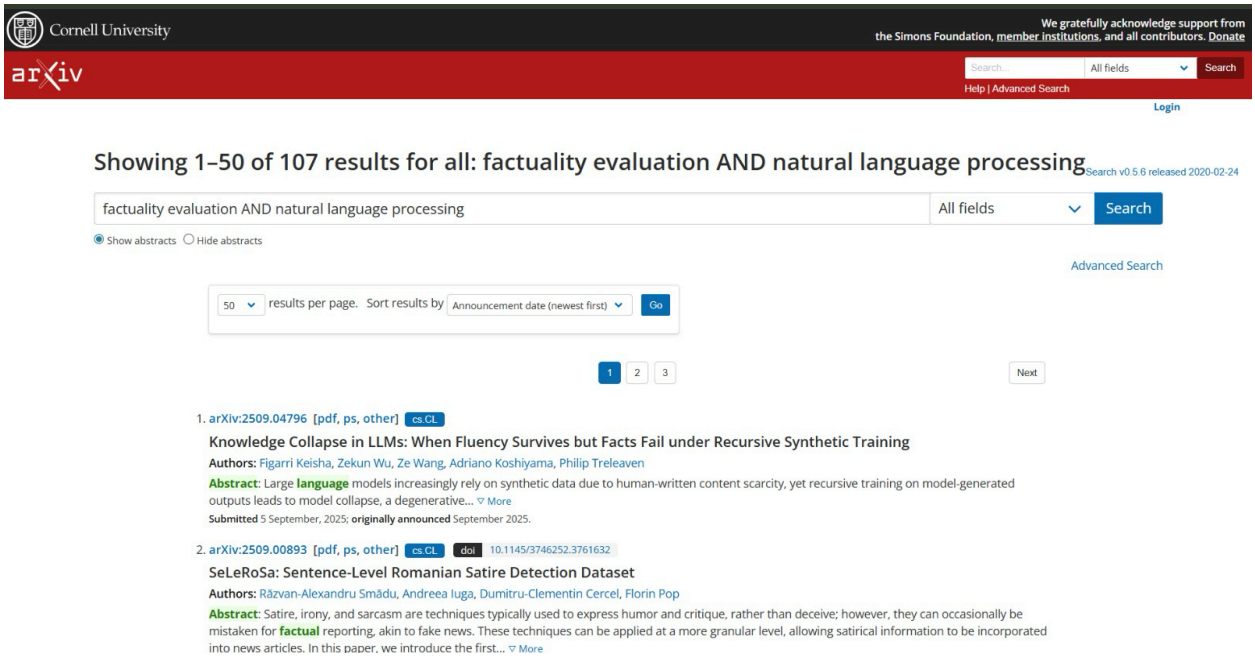
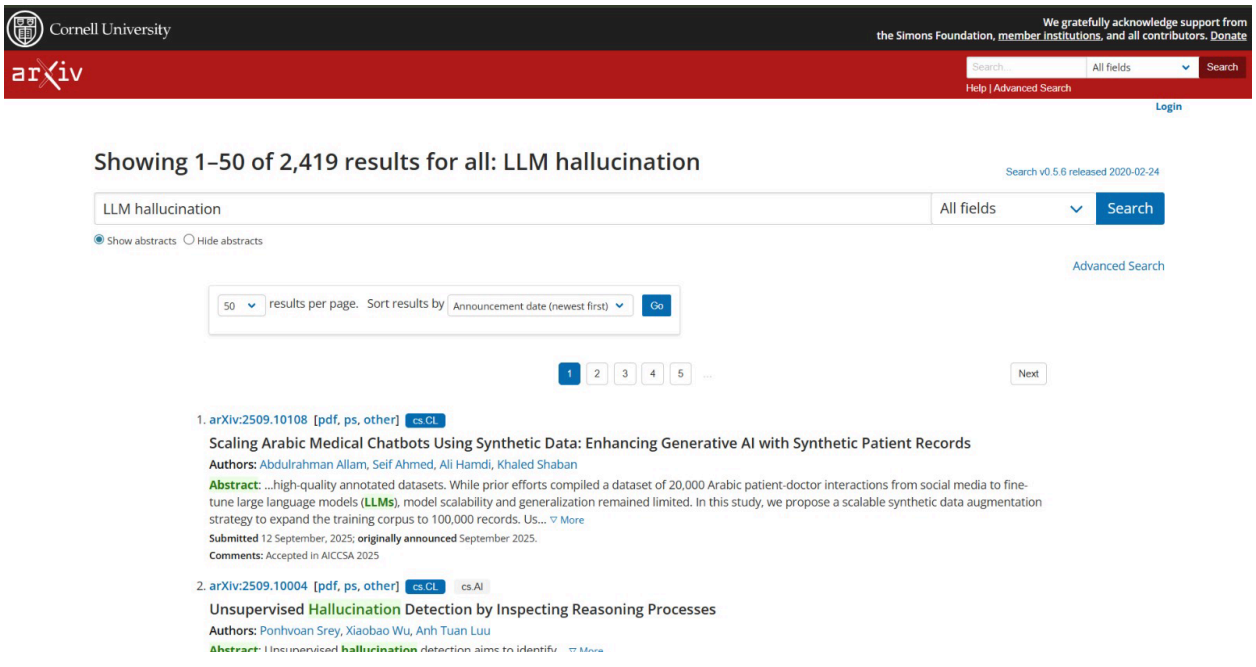
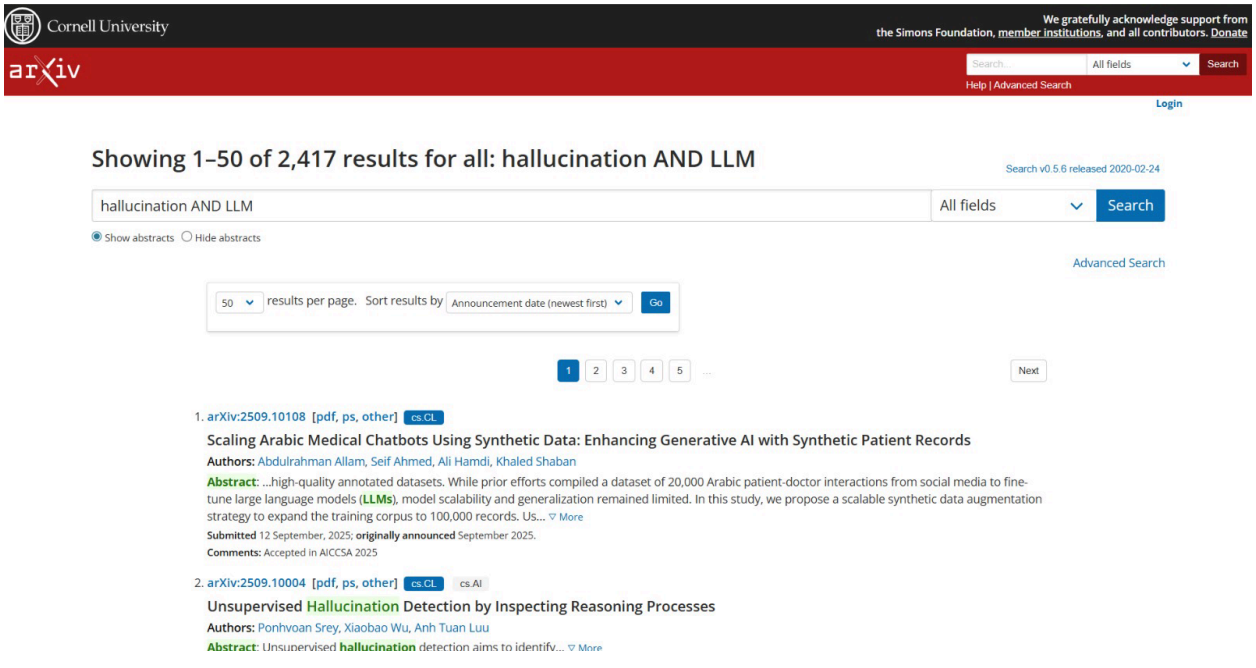

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–25 of 25 results for all: automated fact verification AND LLM
Search v0.5.6 released 2020-02-24
automated fact verification AND LLM
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1. arXiv:2509.08803 [pdf, ps, other] cs.SI cs.AI cs.CL cs.CY
Scaling Truth: The Confidence Paradox in AI Fact-Checking
Authors: Ihsan A. Qazi, Zohaib Khan, Abdullah Ghani, Agha A. Raza, Zafar A. Qazi, Wassay Sajjad, Ayesha Ali, Asher Javaid, Muhammad Abdullah Sohail, Abdul H. Azeemi
Abstract: The rise of misinformation underscores the need for scalable and reliable fact-checking solutions. Large language models (... More
Submitted 10 September, 2025; originally announced September 2025.
Comments: 65 pages, 26 figures, 6 tables
2. arXiv:2508.06495 [pdf, ps, other] cs.CL cs.AI cs.IR
Semi-automated Fact-checking in Portuguese: Corpora Enrichment using Retrieval with Claim extraction
Authors: Juliana Resplande Sant'anna Gomes, Arlindo Rodrigues Galvão Filho
Abstract: The accelerated dissemination of disinformation often outpaces the capacity for manual fact-checking, highlighting the urgent need for Semi-... More
Submitted 19 July, 2025; originally announced August 2025.
Comments: Master Thesis in Computer Science at Federal University on Goias (UFG). Written in Portuguese
3. arXiv:2507.20700 [pdf, ps, other] cs.CL

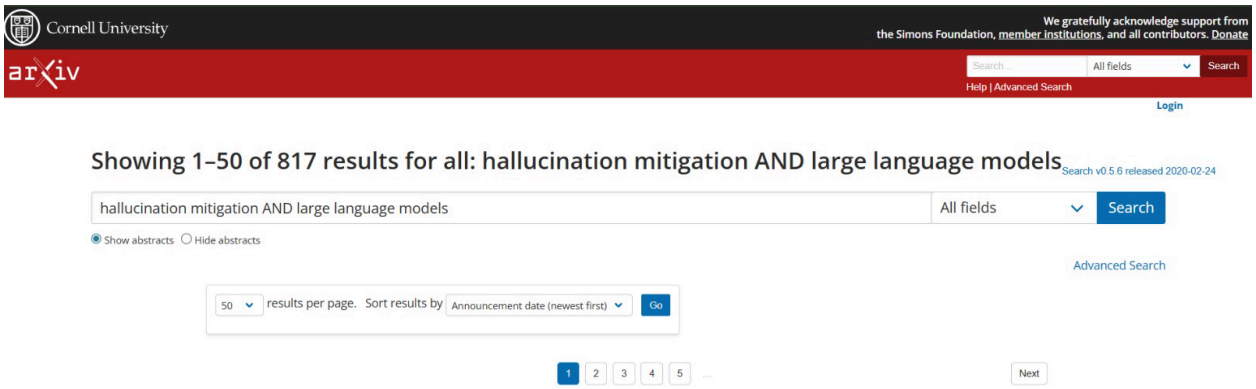

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 283 results for all: large language models AND fact-checking
Search v0.5.6 released 2020-02-24
large language models AND fact-checking
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1 2 3 4 5 6
Next
1. arXiv:2509.09192 [pdf, ps, other] cs.SE cs.AI
Probing Pre-trained Language Models on Code Changes: Insights from ReDef, a High-Confidence Just-in-Time Defect Prediction Dataset
Authors: Doha Nam, Taehyoun Kim, Duksan Ryu, Jongmoon Baik
Abstract: ...in identifying bug-inducing commits. To address this, we present ReDef (Revert-based Defect dataset), a high-confidence benchmark of function-level modifications curated from 22 large-scale C/C++ projects. Defective cases are anchored by revert commits, while clean cases are validated through post-hoc history... More
Submitted 11 September, 2025; originally announced September 2025.
Comments: An anonymous link containing the dataset, construction scripts, and experimental code is publicly available for reproducibility: https://figshare.com/s/4f202bc0921e26b41dc2
2. arXiv:2509.08803 [pdf, ps, other] cs.SI cs.AI cs.CL cs.CY
Scaling Truth: The Confidence Paradox in AI Fact-Checking

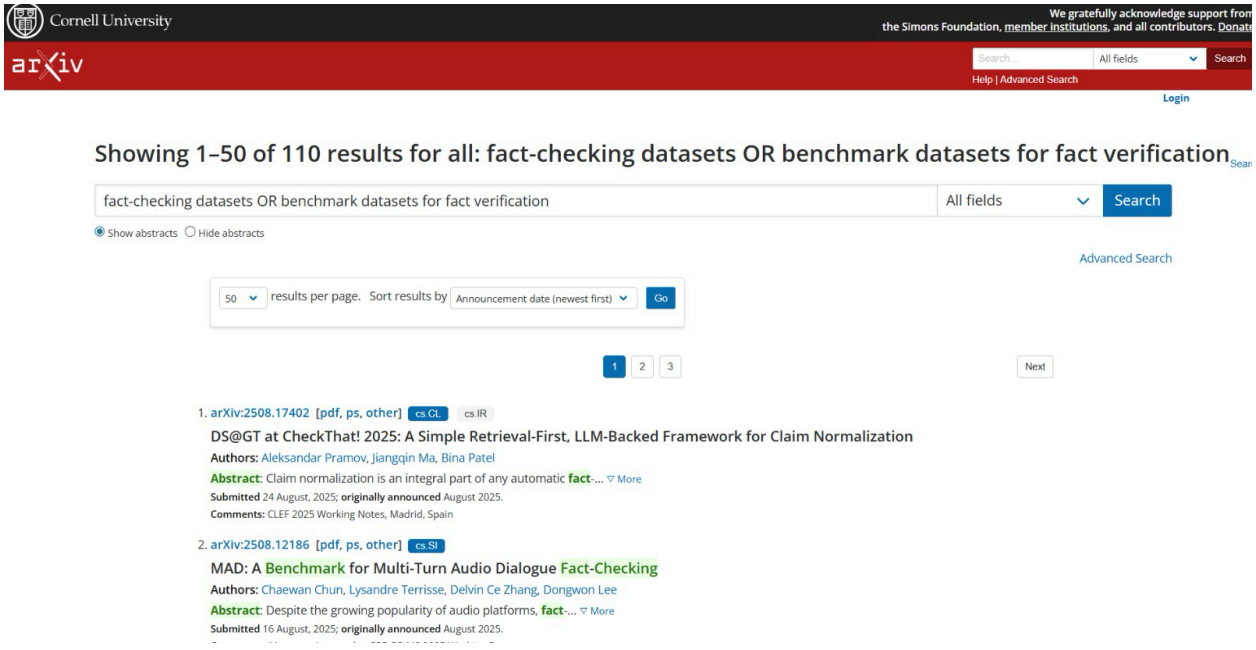

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 165 results for all: LLM AND misinformation detection
Search v0.5.6 released 2020-02-24
LLM AND misinformation detection
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
4
Next
1. arXiv:2508.19633 [pdf, ps, other] cs.CL
A Symbolic Adversarial Learning Framework for Evolving Fake News Generation and Detection
Authors: Chong Tian, Qirong Ho, Xiuying Chen
Abstract: Rapid LLM advancements heighten fake news risks by enabling the automatic generation of increasingly sophisticated... More
Submitted 27 August, 2025; originally announced August 2025.
Comments: Accepted to EMNLP 2025 Main Conference
2. arXiv:2508.18819 [pdf, ps, other] cs.CL cs.SI
LLM-based Contrastive Self-Supervised AMR Learning with Masked Graph Autoencoders for Fake News Detection
Authors: Shubham Gupta, Shraban Kumar Chatterjee, Suman Kundu
Abstract: The proliferation of misinformation in the digital age has led to significant societal challenges. Existing approaches often struggle with capturing long-range dependencies, complex semantic relations, and the social dynamics influencing news dissemination. Furthermore, these methods require extensive labelled datasets, making their deployment resource-inten... More

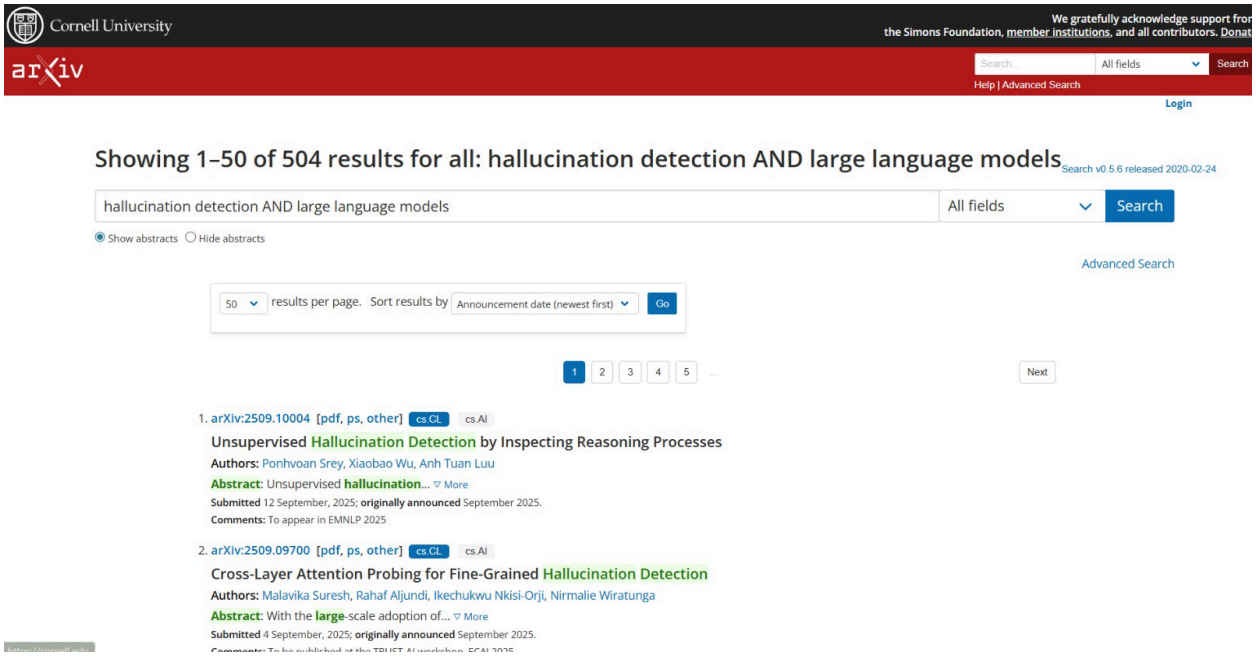

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 107 results for all: factuality evaluation AND natural language processing
Search v0.5.6 released 2020-02-24
factuality evaluation AND natural language processing
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
Next
1. arXiv:2509.04796 [pdf, ps, other] cs.CL
Knowledge Collapse in LLMs: When Fluency Survives but Facts Fail under Recursive Synthetic Training
Authors: Figarri Keisha, Zekun Wu, Ze Wang, Adriano Koshiyama, Philip Treleaven
Abstract: Large language models increasingly rely on synthetic data due to human-written content scarcity, yet recursive training on model-generated outputs leads to model collapse, a degenerative... More
Submitted 5 September, 2025; originally announced September 2025.
2. arXiv:2509.00893 [pdf, ps, other] cs.CL doi 10.1145/3746252.3761632
SeLeRoSa: Sentence-Level Romanian Satire Detection Dataset
Authors: Răzvan-Alexandru Smădu, Andreea Iuga, Dumitru-Clementin Cercel, Florin Pop
Abstract: Satire, irony, and sarcasm are techniques typically used to express humor and critique, rather than deceive; however, they can occasionally be mistaken for factual reporting, akin to fake news. These techniques can be applied at a more granular level, allowing satirical information to be incorporated into news articles. In this paper, we introduce the first... More

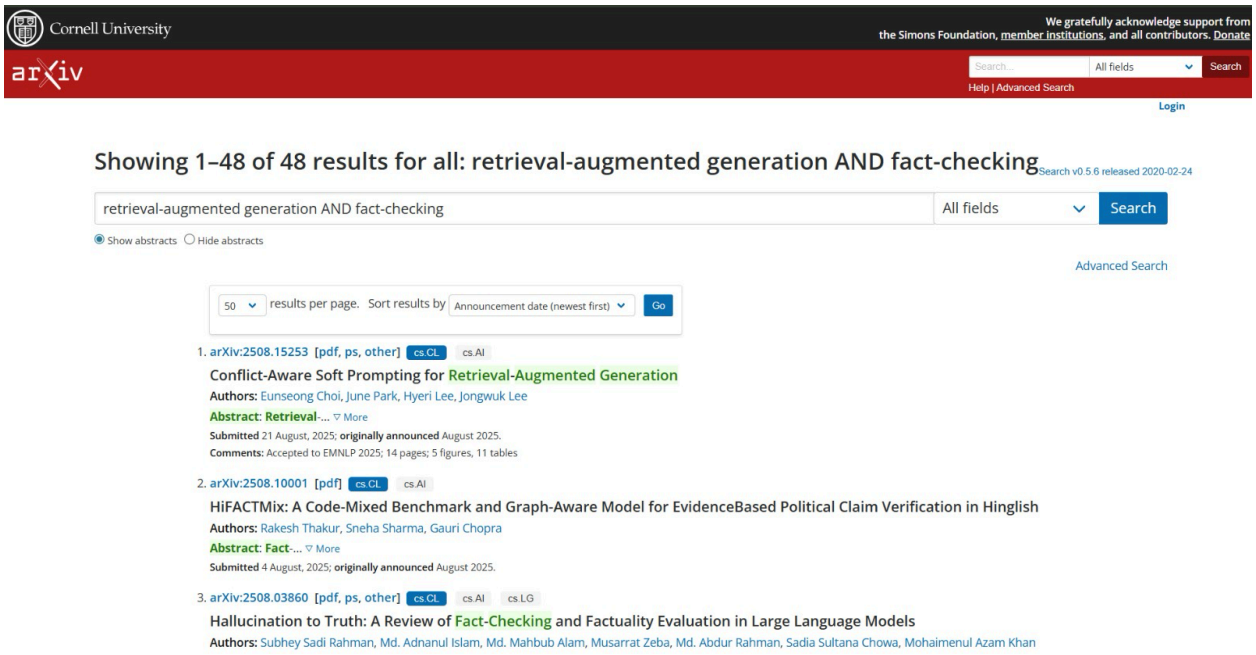

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 2,419 results for all: LLM hallucination
Search v0.5.6 released 2020-02-24
LLM hallucination
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
4
5
Next
1. arXiv:2509.10108 [pdf, ps, other] cs.CL
Scaling Arabic Medical Chatbots Using Synthetic Data: Enhancing Generative AI with Synthetic Patient Records
Authors: Abdulrahman Allam, Seif Ahmed, Ali Hamdi, Khaled Shaban
Abstract: ...high-quality annotated datasets. While prior efforts compiled a dataset of 20,000 Arabic patient-doctor interactions from social media to fine-tune large language models (LLMs), model scalability and generalization remained limited. In this study, we propose a scalable synthetic data augmentation strategy to expand the training corpus to 100,000 records. Us... More
Submitted 12 September, 2025; originally announced September 2025.
Comments: Accepted in AICCSA 2025
2. arXiv:2509.10004 [pdf, ps, other] cs.CL cs.AI
Unsupervised Hallucination Detection by Inspecting Reasoning Processes
Authors: Ponhvoan Srey, Xiaobao Wu, Anh Tuan Luu
Abstract: Unsupervised hallucination detection aims to identify... More

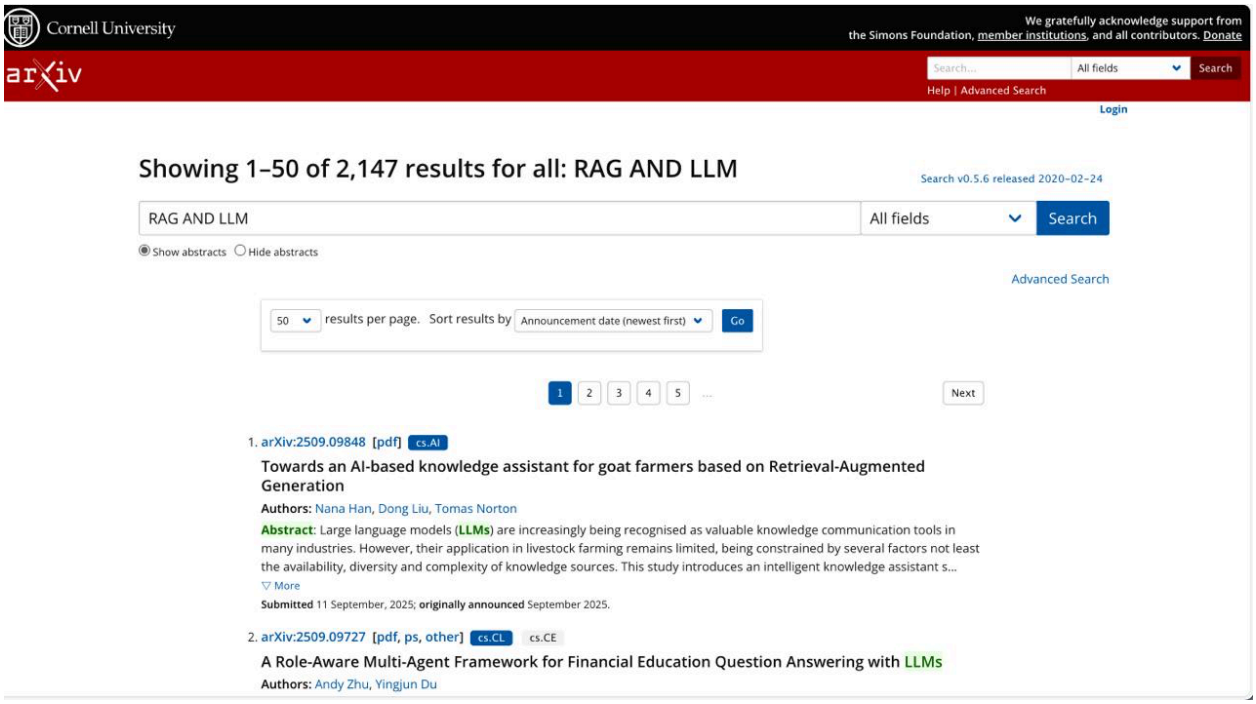

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 2,417 results for all: hallucination AND LLM
Search v0.5.6 released 2020-02-24
hallucination AND LLM
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
4
5
Next
1. arXiv:2509.10108 [pdf, ps, other] cs.CL
Scaling Arabic Medical Chatbots Using Synthetic Data: Enhancing Generative AI with Synthetic Patient Records
Authors: Abdulrahman Allam, Seif Ahmed, Ali Hamdi, Khaled Shaban
Abstract: ...high-quality annotated datasets. While prior efforts compiled a dataset of 20,000 Arabic patient-doctor interactions from social media to fine-tune large language models (LLMs), model scalability and generalization remained limited. In this study, we propose a scalable synthetic data augmentation strategy to expand the training corpus to 100,000 records. Us... More
Submitted 12 September, 2025; originally announced September 2025.
Comments: Accepted in AICCSA 2025
2. arXiv:2509.10004 [pdf, ps, other] cs.CL cs.AI
Unsupervised Hallucination Detection by Inspecting Reasoning Processes
Authors: Ponhvoan Srey, Xiaobao Wu, Anh Tuan Luu
Abstract: Unsupervised hallucination detection aims to identify... More

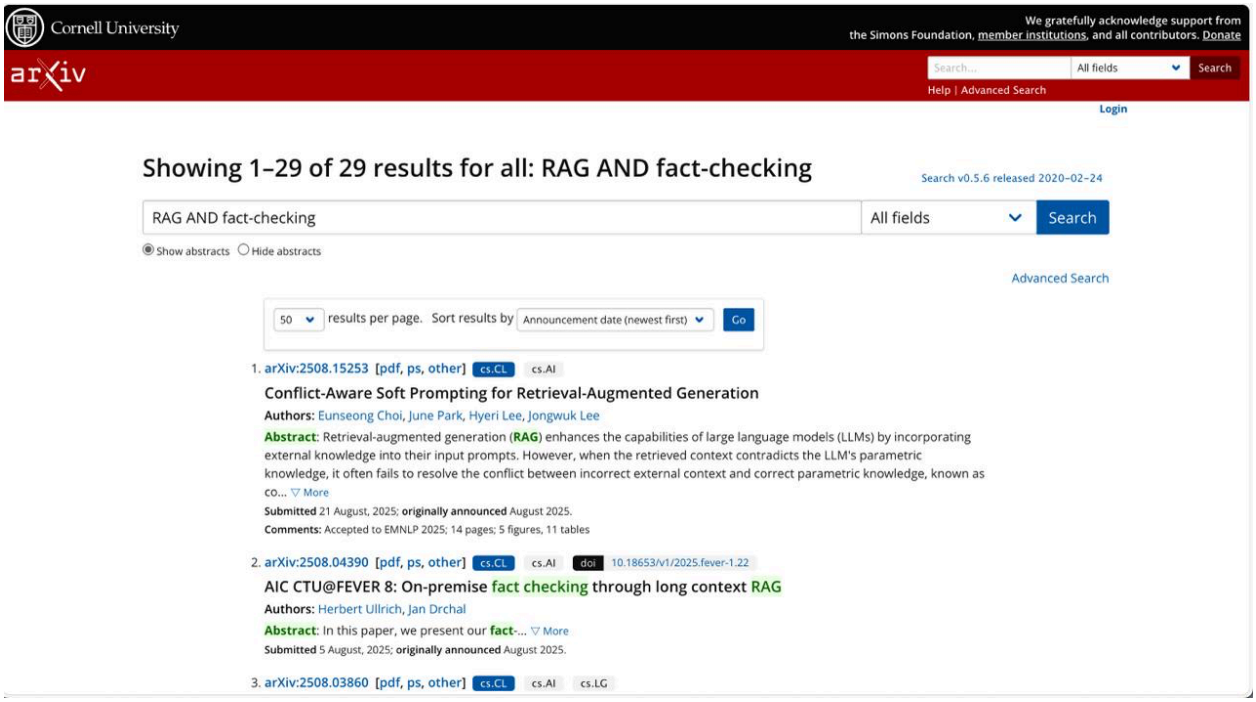

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 817 results for all: hallucination mitigation AND large language models
Search v0.5.6 released 2020-02-24
hallucination mitigation AND large language models
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
4
5
Next
1. arXiv:2509.09700 [pdf, ps, other] cs.CL cs.AI
Cross-Layer Attention Probing for Fine-Grained Hallucination Detection
Authors: Malavika Suresh, Rahaf Aljundi, Ikechukwu Nkisi-Orji, Nirmalie Wiratunga
Abstract: With the large-scale adoption of... More
Submitted 4 September, 2025; originally announced September 2025.
Comments: To be published at the TRUST-AI workshop, ECAI 2025
2. arXiv:2509.07968 [pdf, ps, other] cs.CL
SimpleQA Verified: A Reliable Factuality Benchmark to Measure Parametric Knowledge
Authors: Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, Dipanjan Das
Abstract: We introduce SimpleQA Verified, a 1,000-prompt benchmark for evaluating Large... More
Submitted 9 September, 2025; originally announced September 2025.

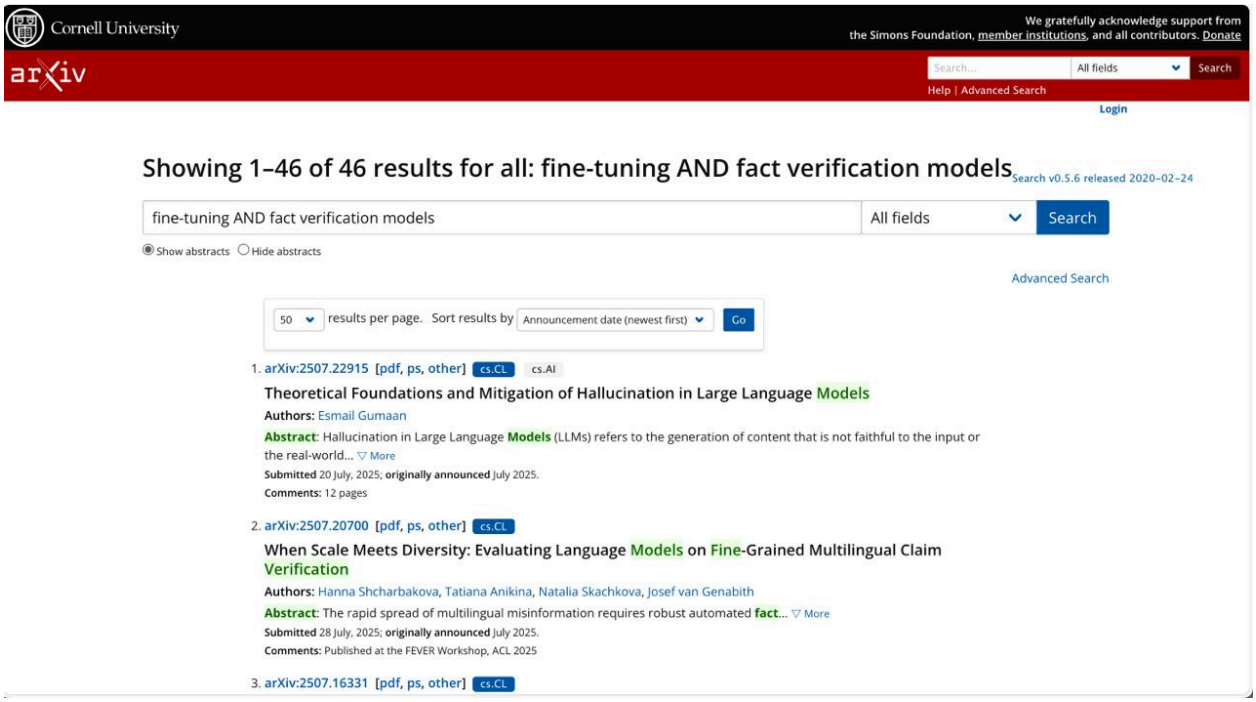

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 110 results for all: fact-checking datasets OR benchmark datasets for fact verification
fact-checking datasets OR benchmark datasets for fact verification
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
Next
1. arXiv:2508.17402 [pdf, ps, other] cs.CL cs.IR
DS@GT at CheckThat! 2025: A Simple Retrieval-First, LLM-Backed Framework for Claim Normalization
Authors: Aleksandar Pramov, Jiangqin Ma, Bina Patel
Abstract: Claim normalization is an integral part of any automatic fact-... More
Submitted 24 August, 2025; originally announced August 2025.
Comments: CLEF 2025 Working Notes, Madrid, Spain
2. arXiv:2508.12186 [pdf, ps, other] cs.SI
MAD: A Benchmark for Multi-Turn Audio Dialogue Fact-Checking
Authors: Chaewan Chun, Lysandre Terrisse, Delvin Ce Zhang, Dongwon Lee
Abstract: Despite the growing popularity of audio platforms, fact-... More
Submitted 16 August, 2025; originally announced August 2025.

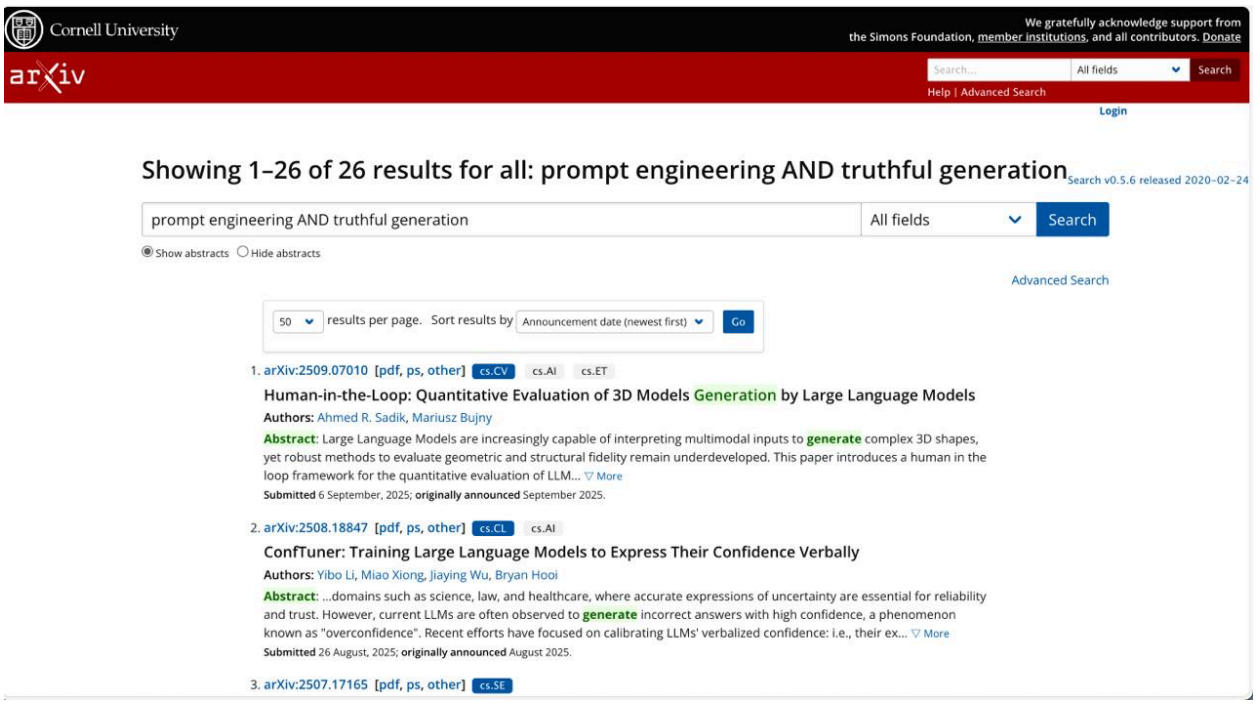

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–50 of 504 results for all: hallucination detection AND large language models
Search v0.5.6 released 2020-02-24
hallucination detection AND large language models
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1
2
3
4
5
Next
1. arXiv:2509.10004 [pdf, ps, other] cs.CL cs.AI
Unsupervised Hallucination Detection by Inspecting Reasoning Processes
Authors: Ponhvoan Srey, Xiaobao Wu, Anh Tuan Luu
Abstract: Unsupervised hallucination... More
Submitted 12 September, 2025; originally announced September 2025.
Comments: To appear in EMNLP 2025
2. arXiv:2509.09700 [pdf, ps, other] cs.CL cs.AI
Cross-Layer Attention Probing for Fine-Grained Hallucination Detection
Authors: Malavika Suresh, Rahaf Aljundi, Ikechukwu Nkisi-Orji, Nirmalie Wiratunga
Abstract: With the large-scale adoption of... More
Submitted 4 September, 2025; originally announced September 2025.

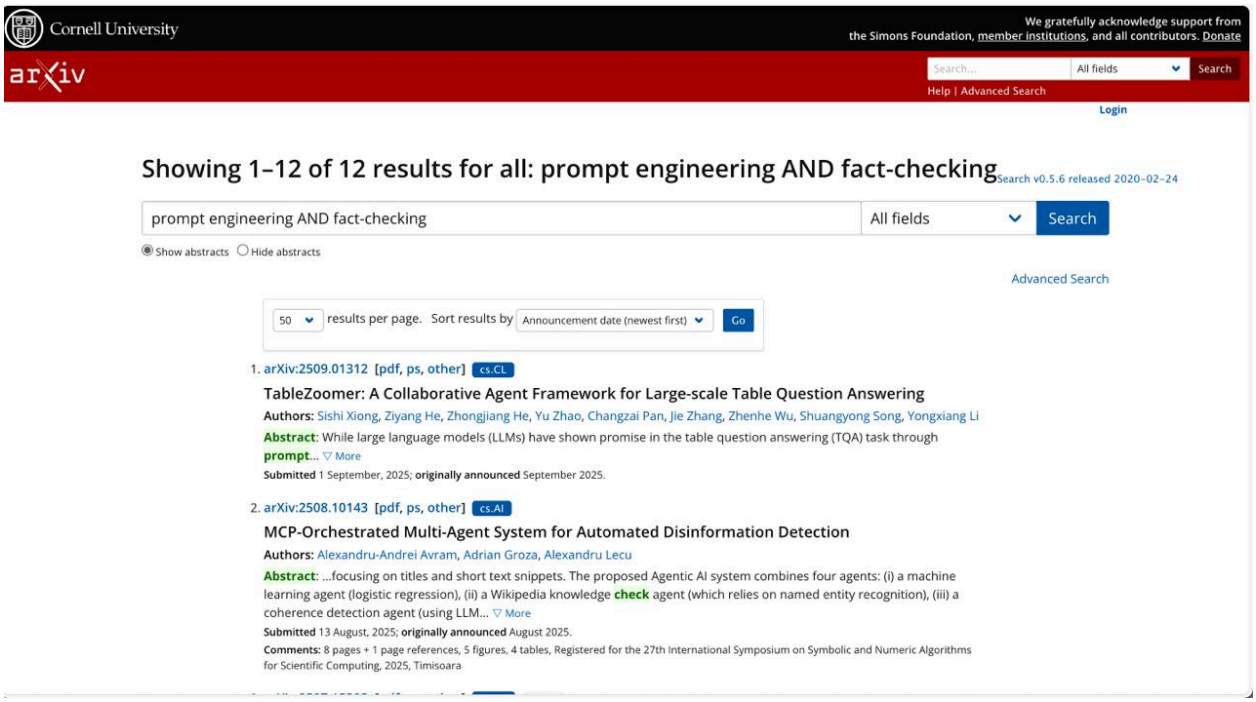

Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–48 of 48 results for all: retrieval-augmented generation AND fact-checking
Search v0.5.6 released 2020-02-24
retrieval-augmented generation AND fact-checking
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1. arXiv:2508.15253 [pdf, ps, other] cs.CL cs.AI
Conflict-Aware Soft Prompting for Retrieval-Augmented Generation
Authors: Eunseong Choi, June Park, Hyeri Lee, Jongwuk Lee
Abstract: Retrieval-... More
Submitted 21 August, 2025; originally announced August 2025.
Comments: Accepted to EMNLP 2025; 14 pages; 5 figures, 11 tables
2. arXiv:2508.10001 [pdf] cs.CL cs.AI
HiFACTMix: A Code-Mixed Benchmark and Graph-Aware Model for EvidenceBased Political Claim Verification in Hinglish
Authors: Rakesh Thakur, Sneha Sharma, Gauri Chopra
Abstract: Fact-... More
Submitted 4 August, 2025; originally announced August 2025.
3. arXiv:2508.03860 [pdf, ps, other] cs.CL cs.AI cs.LG
Hallucination to Truth: A Review of Fact-Checking and Factuality Evaluation in Large Language Models
Authors: Subhey Sadi Rahman, Md. Adnanul Islam, Md. Mahbub Alam, Musarrat Zeba, Md. Abdur Rahman, Sadia Sultana Chowa, Mohaimenul Azam Khan


Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search... All fields Search
Help | Advanced Search
Login
Showing 1–50 of 2,147 results for all: RAG AND LLM
Search v0.5.6 released 2020-02-24
RAG AND LLM All fields Search
Show abstracts Hide abstracts
Advanced Search
50 results per page. Sort results by Announcement date (newest first) Go
1 2 3 4 5 ... Next
1. arXiv:2509.09848 [pdf] cs.AI
Towards an AI-based knowledge assistant for goat farmers based on Retrieval-Augmented Generation
Authors: Nana Han, Dong Liu, Tomas Norton
Abstract: Large language models (LLMs) are increasingly being recognised as valuable knowledge communication tools in many industries. However, their application in livestock farming remains limited, being constrained by several factors not least the availability, diversity and complexity of knowledge sources. This study introduces an intelligent knowledge assistant s...
▽ More
Submitted 11 September, 2025; originally announced September 2025.
2. arXiv:2509.09727 [pdf, ps, other] cs.CL cs.CE
A Role-Aware Multi-Agent Framework for Financial Education Question Answering with LLMs
Authors: Andy Zhu, Yingjun Du


Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search... All fields Search
Help | Advanced Search
Login
Showing 1–29 of 29 results for all: RAG AND fact-checking
Search v0.5.6 released 2020-02-24
RAG AND fact-checking All fields Search
Show abstracts Hide abstracts
Advanced Search
50 results per page. Sort results by Announcement date (newest first) Go
1. arXiv:2508.15253 [pdf, ps, other] cs.CL cs.AI
Conflict-Aware Soft Prompting for Retrieval-Augmented Generation
Authors: Eunseong Choi, June Park, Hyeri Lee, Jongwuk Lee
Abstract: Retrieval-augmented generation (RAG) enhances the capabilities of large language models (LLMs) by incorporating external knowledge into their input prompts. However, when the retrieved context contradicts the LLM's parametric knowledge, it often fails to resolve the conflict between incorrect external context and correct parametric knowledge, known as co... ▽ More
Submitted 21 August, 2025; originally announced August 2025.
Comments: Accepted to EMNLP 2025; 14 pages; 5 figures, 11 tables
2. arXiv:2508.04390 [pdf, ps, other] cs.CL cs.AI doi 10.18653/v1/2025.fever-1.22
AIC CTU@FEVER 8: On-premise fact checking through long context RAG
Authors: Herbert Ullrich, Jan Drchal
Abstract: In this paper, we present our fact-... ▽ More
Submitted 5 August, 2025; originally announced August 2025.
3. arXiv:2508.03860 [pdf, ps, other] cs.CL cs.AI cs.LG


Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search... All fields Search
Help | Advanced Search
Login
Showing 1–46 of 46 results for all: fine-tuning AND fact verification models
Search v0.5.6 released 2020-02-24
fine-tuning AND fact verification models
All fields
Search
Show abstracts Hide abstracts
Advanced Search
50 results per page. Sort results by Announcement date (newest first) Go
1. arXiv:2507.22915 [pdf, ps, other] cs.CL cs.AI
Theoretical Foundations and Mitigation of Hallucination in Large Language Models
Authors: Esmail Gumaan
Abstract: Hallucination in Large Language Models (LLMs) refers to the generation of content that is not faithful to the input or the real-world... More
Submitted 20 July, 2025; originally announced July 2025.
Comments: 12 pages
2. arXiv:2507.20700 [pdf, ps, other] cs.CL
When Scale Meets Diversity: Evaluating Language Models on Fine-Grained Multilingual Claim Verification
Authors: Hanna Shcharbakova, Tatiana Anikina, Natalia Skachkova, Josef van Genabith
Abstract: The rapid spread of multilingual misinformation requires robust automated fact... More
Submitted 28 July, 2025; originally announced July 2025.
Comments: Published at the FEVER Workshop, ACL 2025
3. arXiv:2507.16331 [pdf, ps, other] cs.CL


Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search... All fields Search
Help | Advanced Search
Login
Showing 1–26 of 26 results for all: prompt engineering AND truthful generation
Search v0.5.6 released 2020-02-24
prompt engineering AND truthful generation
All fields
Search
Show abstracts Hide abstracts
Advanced Search
50 results per page. Sort results by Announcement date (newest first) Go
1. arXiv:2509.07010 [pdf, ps, other] cs.CV cs.AI cs.ET
Human-in-the-Loop: Quantitative Evaluation of 3D Models Generation by Large Language Models
Authors: Ahmed R. Sadik, Mariusz Bujny
Abstract: Large Language Models are increasingly capable of interpreting multimodal inputs to generate complex 3D shapes, yet robust methods to evaluate geometric and structural fidelity remain underdeveloped. This paper introduces a human in the loop framework for the quantitative evaluation of LLM... More
Submitted 6 September, 2025; originally announced September 2025.
2. arXiv:2508.18847 [pdf, ps, other] cs.CL cs.AI
ConfTuner: Training Large Language Models to Express Their Confidence Verbally
Authors: Yibo Li, Miao Xiong, Jiaying Wu, Bryan Hooi
Abstract: ...domains such as science, law, and healthcare, where accurate expressions of uncertainty are essential for reliability and trust. However, current LLMs are often observed to generate incorrect answers with high confidence, a phenomenon known as "overconfidence". Recent efforts have focused on calibrating LLMs' verbalized confidence: i.e., their ex... More
Submitted 26 August, 2025; originally announced August 2025.
3. arXiv:2507.17165 [pdf, ps, other] cs.SE



Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv

Search... All fields Search

Help | Advanced Search

Login

Showing 1–12 of 12 results for all: prompt engineering AND fact-checking

Search v0.5.6 released 2020-02-24

prompt engineering AND fact-checking All fields Search

Show abstracts Hide abstracts

Advanced Search

50 results per page. Sort results by Announcement date (newest first) Go

1. arXiv:2509.01312 [pdf, ps, other] cs.CL

TableZoomer: A Collaborative Agent Framework for Large-scale Table Question Answering

Authors: Sishi Xiong, Ziyang He, Zhongjiang He, Yu Zhao, Changzai Pan, Jie Zhang, Zhenhe Wu, Shuangyong Song, Yongxiang Li

Abstract: While large language models (LLMs) have shown promise in the table question answering (TQA) task through **prompt**... ▽ More

Submitted 1 September, 2025; **originally announced** September 2025.

2. arXiv:2508.10143 [pdf, ps, other] cs.AI

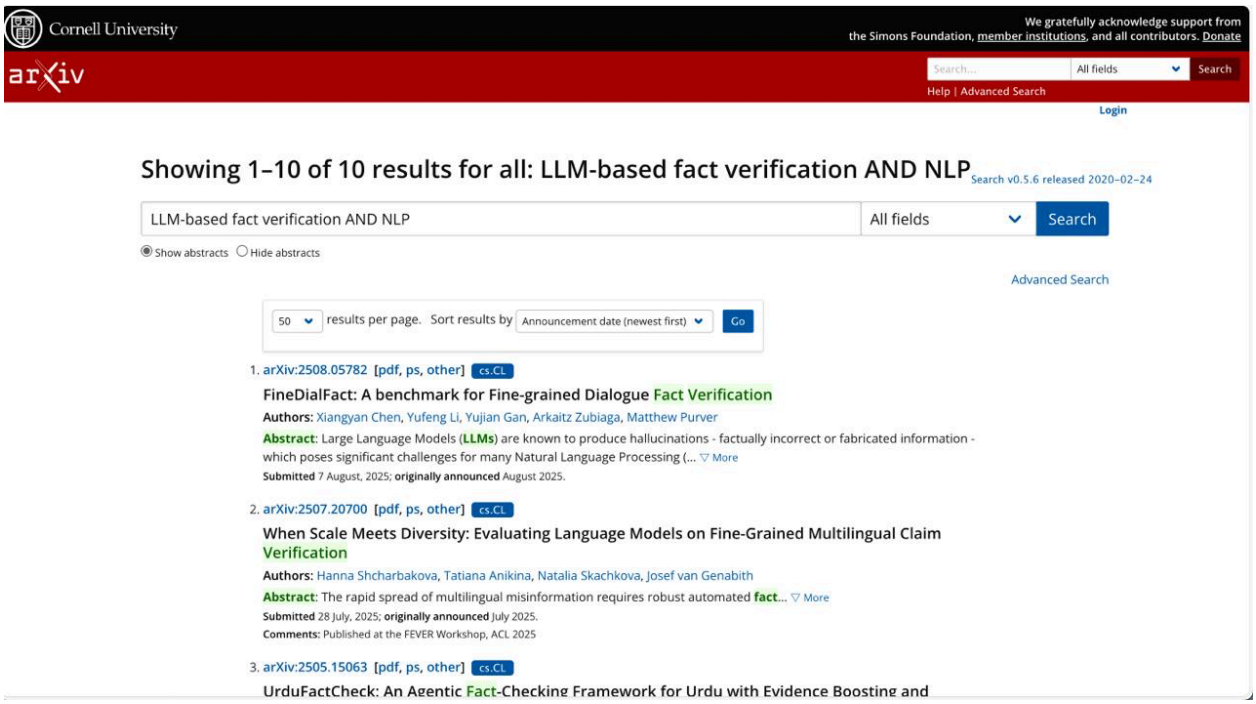
MCP-Orchestrated Multi-Agent System for Automated Disinformation Detection

Authors: Alexandru-Andrei Avram, Adrian Groza, Alexandru Lecu

Abstract: ...focusing on titles and short text snippets. The proposed Agentic AI system combines four agents: (i) a machine learning agent (logistic regression), (ii) a Wikipedia knowledge **check** agent (which relies on named entity recognition), (iii) a coherence detection agent (using LLM... ▽ More

Submitted 13 August, 2025; **originally announced** August 2025.

Comments: 8 pages + 1 page references, 5 figures, 4 tables, Registered for the 27th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2025, Timisoara



Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv

Search... All fields Search

Help | Advanced Search

Login

Showing 1–10 of 10 results for all: LLM-based fact verification AND NLP

Search v0.5.6 released 2020-02-24

LLM-based fact verification AND NLP All fields Search

Show abstracts Hide abstracts

Advanced Search

50 results per page. Sort results by Announcement date (newest first) Go

1. arXiv:2508.05782 [pdf, ps, other] cs.CL

FineDialFact: A benchmark for Fine-grained Dialogue Fact Verification

Authors: Xiangyan Chen, Yufeng Li, Yujian Gan, Arkaitz Zubiaga, Matthew Purver

Abstract: Large Language Models (**LLMs**) are known to produce hallucinations - factually incorrect or fabricated information - which poses significant challenges for many Natural Language Processing (... ▽ More

Submitted 7 August, 2025; **originally announced** August 2025.

2. arXiv:2507.20700 [pdf, ps, other] cs.CL

When Scale Meets Diversity: Evaluating Language Models on Fine-Grained Multilingual Claim Verification

Authors: Hanna Shcharbakova, Tatiana Anikina, Natalia Skachkova, Josef van Genabith

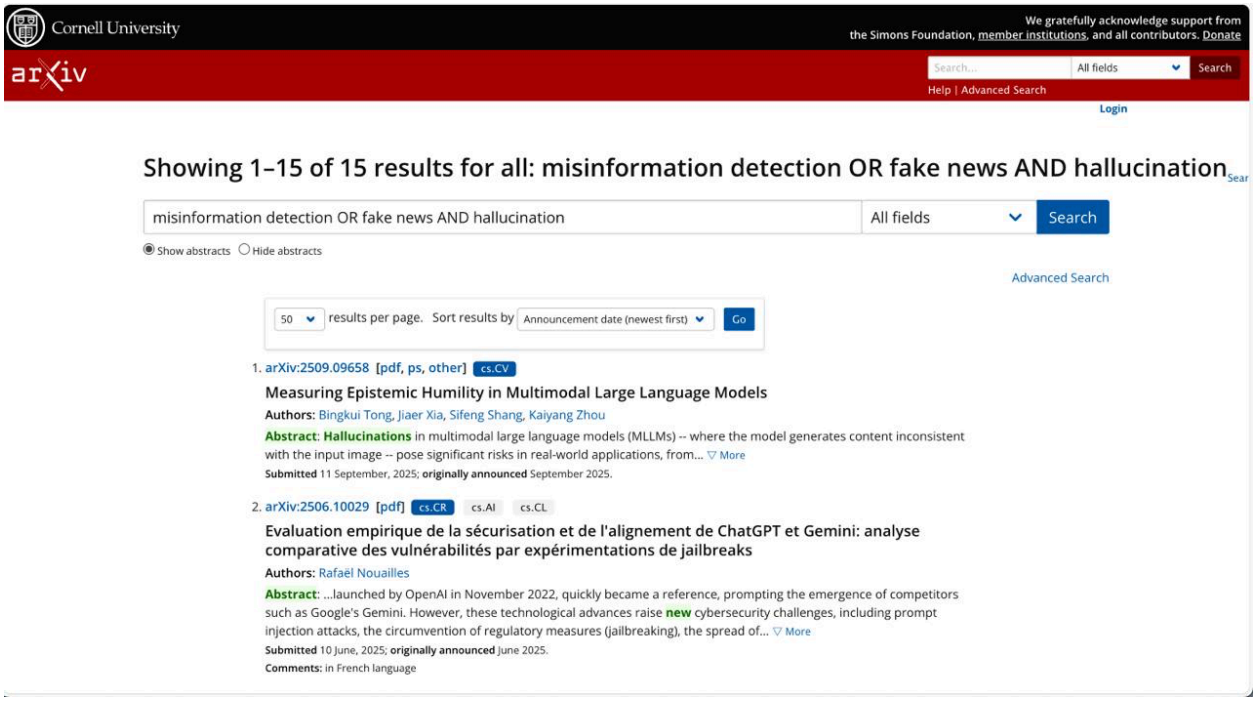
Abstract: The rapid spread of multilingual misinformation requires robust automated **fact**... ▽ More

Submitted 28 July, 2025; **originally announced** July 2025.

Comments: Published at the FEVER Workshop, ACL 2025

3. arXiv:2505.15063 [pdf, ps, other] cs.CL

UrduFactCheck: An Agentic Fact-Checking Framework for Urdu with Evidence Boosting and


Cornell University
We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate
arXiv
Search...
All fields
Search
Help | Advanced Search
Login
Showing 1–15 of 15 results for all: misinformation detection OR fake news AND hallucination
misinformation detection OR fake news AND hallucination
All fields
Search
Show abstracts
Hide abstracts
Advanced Search
50
results per page.
Sort results by
Announcement date (newest first)
Go
1. arXiv:2509.09658 [pdf, ps, other] cs.CV
Measuring Epistemic Humility in Multimodal Large Language Models
Authors: Bingkui Tong, Jiaer Xia, Sifeng Shang, Kaiyang Zhou
Abstract: Hallucinations in multimodal large language models (MLLMs) -- where the model generates content inconsistent with the input image -- pose significant risks in real-world applications, from... ▽ More
Submitted 11 September, 2025; originally announced September 2025.
2. arXiv:2506.10029 [pdf] cs.CR cs.AI cs.CL
Evaluation empirique de la sécurisation et de l'alignement de ChatGPT et Gemini: analyse comparative des vulnérabilités par expérimentations de jailbreaks
Authors: Rafaël Nouailles
Abstract: ...launched by OpenAI in November 2022, quickly became a reference, prompting the emergence of competitors such as Google's Gemini. However, these technological advances raise new cybersecurity challenges, including prompt injection attacks, the circumvention of regulatory measures (jailbreaking), the spread of... ▽ More
Submitted 10 June, 2025; originally announced June 2025.
Comments: in French language