

Bias in LLMs: Measurement, Control, and Risks

Prof. Gianluca Demartini

Australian Research Council Future Fellow at the School of Electrical Engineering and Computer Science – UQ

Dieter Schwarz Fellow at the Institute for Advanced Study – TUM

Research Interests

- **Information Access** (since 2005)

Structured/Unstructured data (SIGIR12), Entity Types (ISWC13, WSemJ16)
Entity Recognition (WWW14, CSUR), Prepositions (CIKM14), Entity Cards (SIGIR19)
Eval (ECIR16 BestP, CIKM17, SIGIR18, CIKM19, WWW22, TOIS23, ICTIR23 BestP, ECIR26)

- **Human-AI Systems** (since 2012)

Entity Linking (WWW12, VLDBJ), CrowdQ (CIDR13), Learnersourcing (LAK21, LAK22, JCAL)
LLM (COLING25, CHI25), Misinfo (ECIR20 Best SP, SIGIR20, CIKM20, IP&M, ICWSM24)
Decision making (CHB-R25, HBET26), Critical Thinking (IP&M25)

- **Better Crowdsourcing Platforms** (since 2013)

Platforms (WWW15, CSCWJ18, CACM25), Experiments (CSCW21), Pricing (HCOMP14)
Task Allocation (WWW13, WWW16, COR), Workers (CHI15, CSCW20 Hon. Mention)
Metadata (IP&M), Attacks (HCOMP18 Best P, JAIR), Time (HCOMP16)
Modus Operandi (UBICOMP17, HT19, WSDM20, TOIS24), Complexity (HCOMP16)
Abandonment (WSDM19, TKDE, ACM TSC)

- **Data Bias and Better Data** (since 2018)

Gender (SIGIR18, ACIS24, WWW25), Management (CACM24), SES (WebSci22, ICWSM25)
Impact on ML (CIKM22), Political (WWW25, ICWSM26, TIST), Difficulty (ACL26)
Noise (WWW19, ICWSM25), Data Workers (SIGIR20, TOIS, TKDE, WWW23, EMNLP26)
Know. Graphs (ISWC19), Unknown Unknowns (ECAI20, HCOMP21), Behaviors (CIKM20)
Fairness (CIKM22, SIGIR23, FAccT24, KDD24, TMLR), Active Learning (AAAI24)

Thanks to:



Australian Government
Australian Research Council



Swiss National
Science Foundation



Engineering and
Physical Sciences
Research Council



My life after L3S

L3S 2006-2011

A/Prof UQ
2019-2024

Running

Post-doc in
Switzerland
2011-2014

Lecturer, Sheffield
2014-2017

S. Lecturer UQ
2017-2019



20y L3S in 2022



Turtles



Full Prof UQ
2025-now



UC Berkeley 2012

Kid 2017-now



HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

Stable Diffusion's text-to-image model amplifies stereotypes about race and gender – here's why that matters

Use of AI could worsen racism and sexism in Australia, human rights commissioner warns

The Problem – Bias in AI



United Nations

UN News

Global perspective Human stories

Artificial intelligence: rooting out bias and stereotypes



Guardian

[Business](#)
[Health](#)
[Science](#)

THE CONVERSATION

Academic rigour, journalistic flair

Search analysis, research, a

[Arts + Culture](#)
[Books + Ideas](#)
[Business + Economy](#)
[Education](#)
[Environment + Energy](#)
[Health](#)
[Politics + Society](#)
[Science + Tech](#)



Historical images made with AI recycle colonial stereotypes and bias – new research

Published: October 24, 2025 9:54am AEDT

Sport

Culture

Lifestyle



The Guardian

Aus

[Environment](#)
[Climate crisis](#)
[Indigenous Australia](#)
[Immigration](#)
[Media](#)
[Business](#)
[Health](#)
[Science](#)
[Tech](#)
[Podcasts](#)
[Newsletters](#)

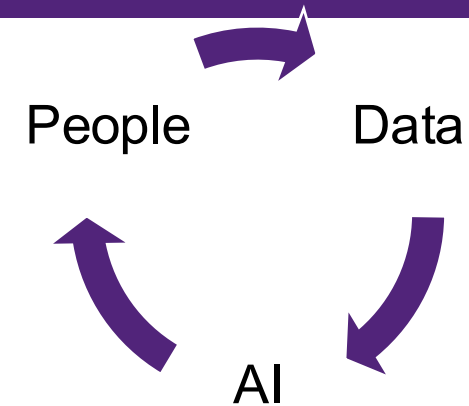
Police AI chief admits crime-fighting tech will have bias but vows to tackle it

Most viewed

Grace Tame says 's

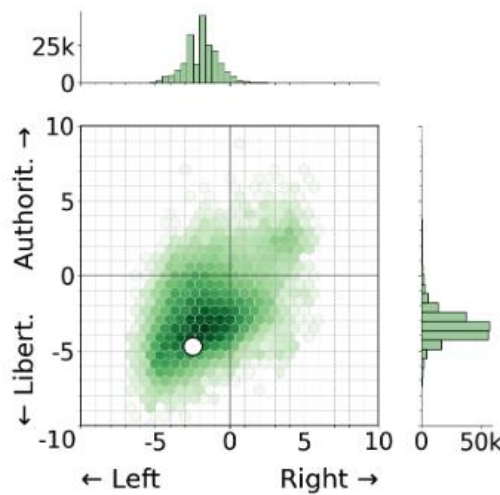
Outline

- Politically biased LLMs doing hate speech detection (IP&M, ACM TIST)
- Bias management (CACM)
- Helping humans spot bias (visualizing bias for better bias annotations) (IP&M)
- The impact of bias definitions (EMNLP 2026)
- Helping criminal investigators solve murder cases (COLING 2025)
- Understanding LLM Task difficulty for routing (ACL 2026)

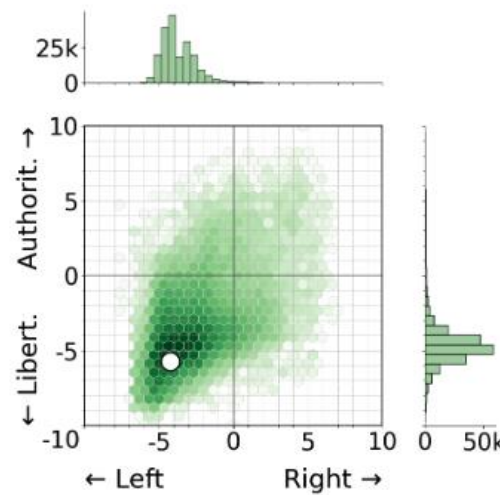


Persona-based LLMs

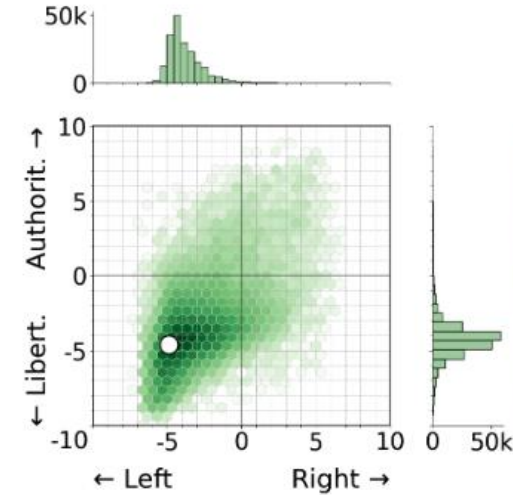
- We make LLMs answer the Political Compass Test
- We then make them impersonate 200,000 personas and answer the PCT again
- This shows how we can measure and control the political bias of LLMs.
- It also highlights embedded stereotypes like: “A business developer trying to bring new investments to the region, regardless of environmental cost” being authoritarian right



Qwen2.5-72B

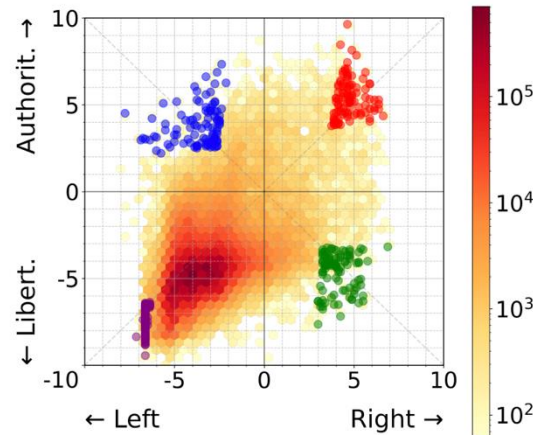
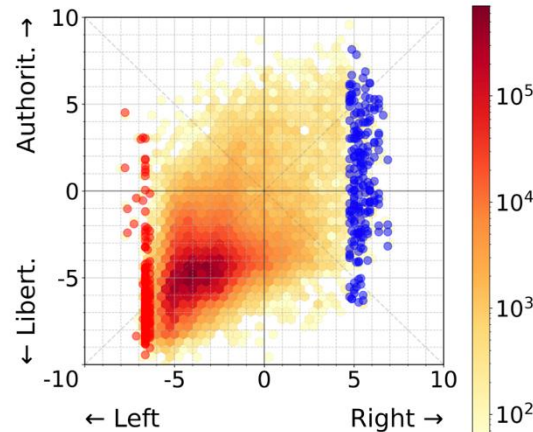


Llama-3.1-70B



Llama-3.3-70B

Harmful content moderation with biased LLMs



(c) Llama-3.1-70B

Target Category	Llama-3.1-8B			Qwen-2.5-32B			Llama-3.1-70B		
	Left	Right	OR	Left	Right	OR	Left	Right	OR
liberals	0.664	0.479	2.141*	0.540	0.373	1.978*	0.696	0.510	2.194*
communists	0.650	0.430	2.466*	0.511	0.390	1.632*	0.688	0.550	1.778*
democrats	0.548	0.415	1.709*	0.498	0.444	1.240*	0.649	0.490	1.925*
left-wingers	0.605	0.420	2.116*	0.484	0.351	1.736*	0.612	0.414	2.233*
right-wingers	0.566	0.425	1.761*	0.441	0.504	0.777*	0.631	0.759	0.543*
republicans	0.512	0.406	1.536*	0.451	0.514	0.776*	0.555	0.680	0.589*
conservatives	0.562	0.420	1.774*	0.434	0.509	0.740*	0.539	0.616	0.728*
Overall	0.601	0.427	2.021	0.487	0.396	1.445	0.628	0.505	1.652

Hate speech detection rates by target category and persona position

Stefano Civelli, Pietro Bernardelle, Nardiana A. Pratama, and Gianluca Demartini. **Ideology-Based LLMs for Content Moderation**. In: ACM Transactions on Intelligent Systems and Technology (TIST). 2026.

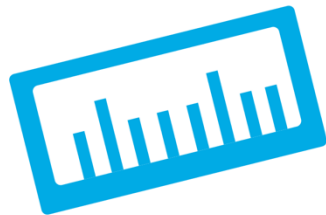
Bias Management, not bias removal

Employing an explicit and not transparent bias removal intervention might be potentially harmful to the user

Figure 2. The five steps of bias management.



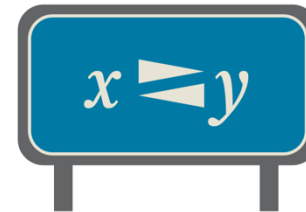
Identifying



Measuring



Indexing



Surfacing



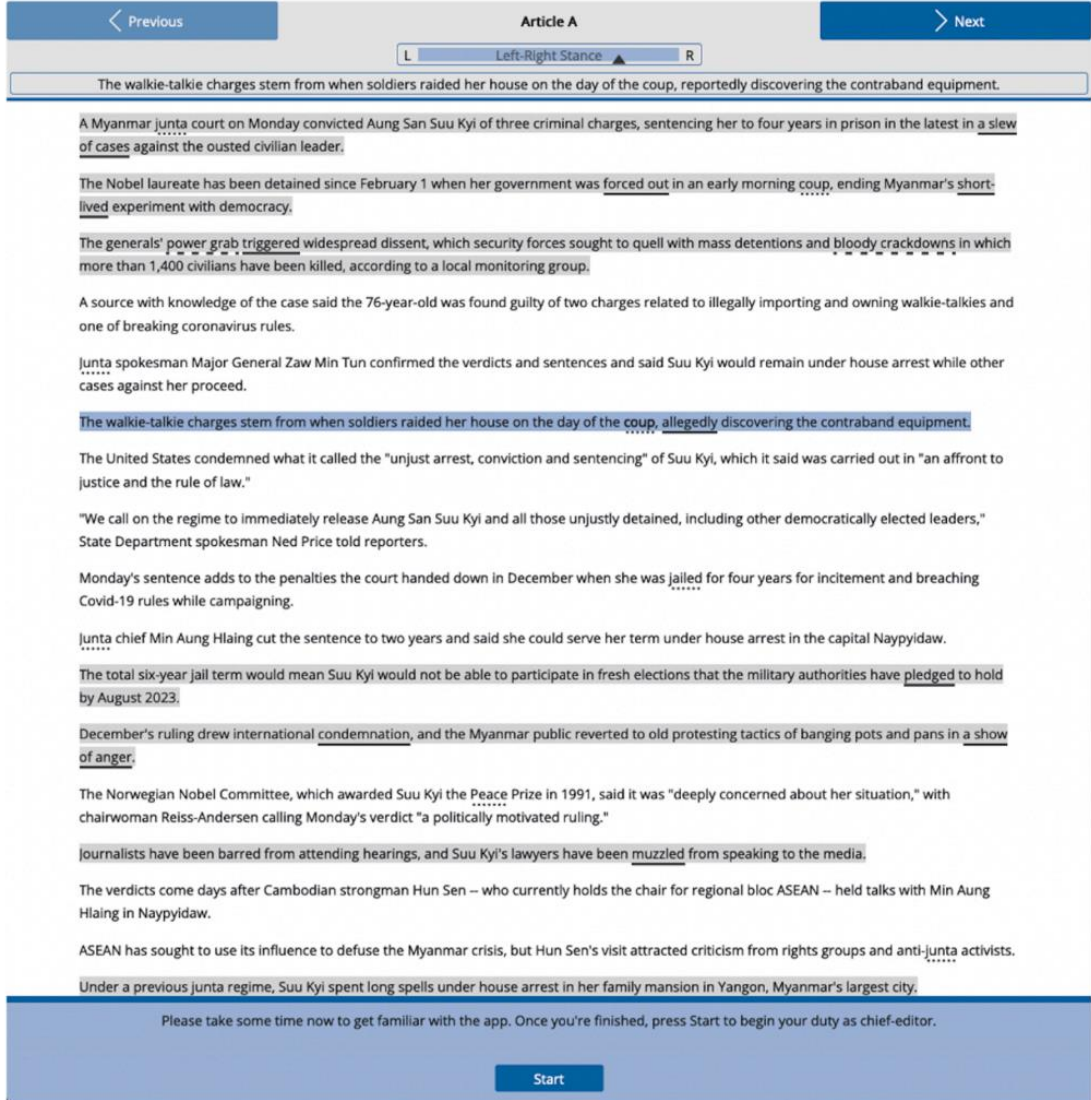
Adapting

Helping humans to spot bias

Testing the role of AI-generated bias labels in training people to spot bias

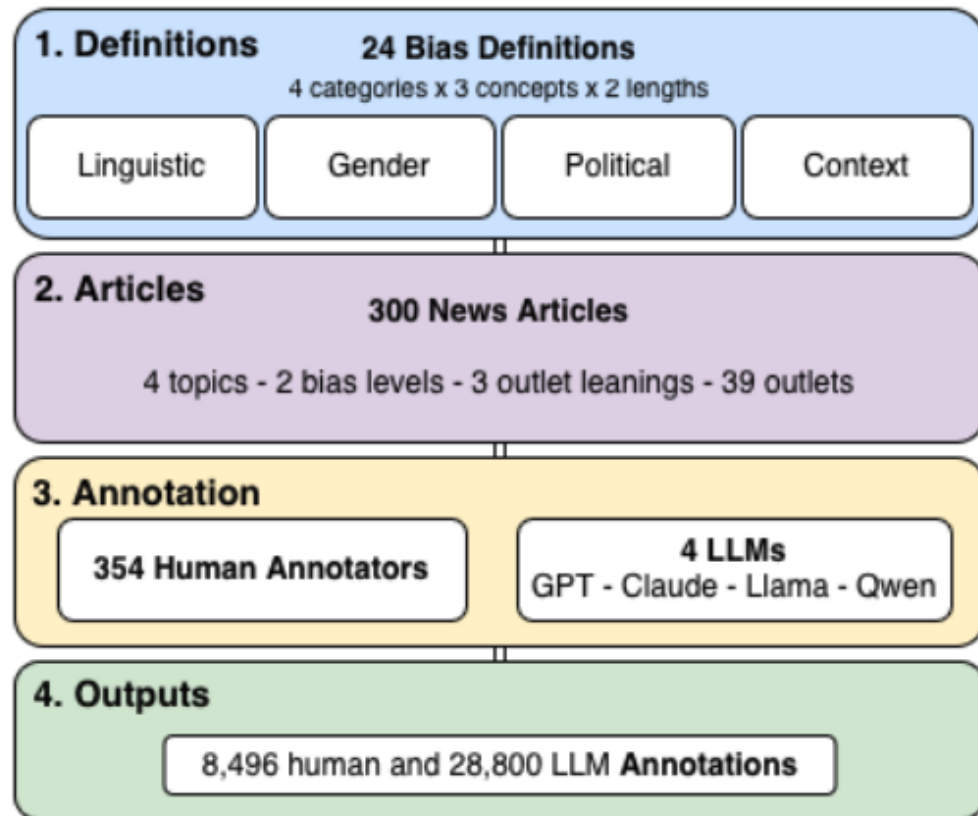
Two phases: i) train subjects with human/ai/no bias annotations and then ii) ask them to spot biased statements

Marking biased texts effectively increases media bias awareness



Timo Spinde, Fei Wu, Wolfgang Gaissmaier, Gianluca Demartini, and Helge Giese. **Enhancing Media Literacy: The Effectiveness of (Human) Annotations and Bias Visualizations on Bias Detection.** In: Information Processing & Management (IP&M), Elsevier, June 2025.

Spotting bias: Humans vs LLMs



Observations:

- Definitions consistently affects human and LLM ratings
- Definitions matter most where annotation is already difficult
- LLMs exhibit larger definition effects than humans (i.e., they follow instructions better than humans)
- LLMs rate bias systematically lower than humans

Martin Paul Wessel, Timo Spinde, Jürgen Pfeffer, and Gianluca Demartini. **Definitional Sensitivity in Media Bias Detection: A Multi-Definition Dataset and Benchmark.** In: Findings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026). Budapest, Hungary, October 2026.

RAG for criminal investigation – fabricated information

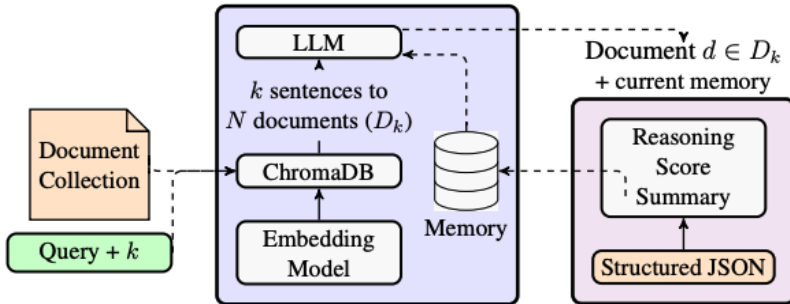


Figure 1: From a document collection and queries, Kri-RAG finds similar sentences and analyzes connected documents. A memory is built for each query as it loops through the documents.

Metric	Model	Q1	Q2	Q3	Q4
P	Gemma-2-9B	0.43	0.50	0.80	0.33
	Gemma-2-27b	1.00	0.44	0.88	0.50
	Llama-3.1-8b	0.67	0.50	0.62	0.17
	Phi-3-Medium	0.12	0.38	0.71	0.00
R	Gemma-2-9B	0.60	0.29	0.24	0.40
	Gemma-2-27b	0.40	0.24	0.41	0.40
	Llama-3.1-8b	0.40	0.29	0.47	0.40
	Phi-3-Medium	0.20	0.18	0.29	0.00
F_1	Gemma-2-9B	0.50	0.37	0.36	0.36
	Gemma-2-27b	0.57	0.31	0.56	0.44
	Llama-3.1-8b	0.50	0.37	0.53	0.24
	Phi-3-Medium	0.15	0.24	0.42	0.00

Type of misinformation	Occurrences	Example output	Explanation
Fabrication of a single detail	76/117	Stabbed with a knife	Document does not contain 'knife'
Fabrication of several details	13/117	He has a history of violent altercations, including a bar fight and a road rage incident.	Document does not mention these details
Contextual error	6/117	Document 10 is an autopsy report and would directly address the cause of death.	Document 10 is not an autopsy
Misinterpretation	22/117	Between 22:10 and 23:10.	Document mentions dates, not time.

Mads Skipanes, Tollef Emil Jørgensen, Kyle Porter, Gianluca Demartini, and Sule Yildirim Yayilgan. **Enhancing Criminal Investigation Analysis with Summarization and Memory-based Retrieval-Augmented Generation: A Comprehensive Evaluation of Real Case Data.** In: The 31st International Conference on Computational Linguistics (COLING 2025). Abu Dhabi, January 2025.

Routing prompts to one LLM among several options

A way to manage bias is to decide which LLM to send a prompt to and get an answer from

- For example, you have 200k politically-biased LLMs

This may allow users to:

- Pick the most appropriate LLM given their bias
- Save money by using cheaper LLMs for **simpler questions** and powerful LLMs for **complex questions**

Estimating the difficulty of a prompt

To perform routing we need to be able to estimate the complexity of a question.

We use internal layers and linear probes to **estimate the difficulty of questions** given to an LLM.

Middle layers give us the best estimation.

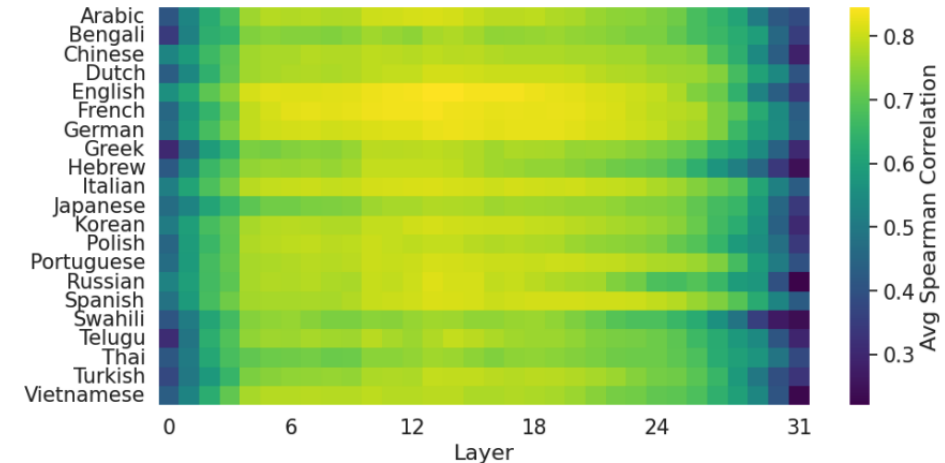
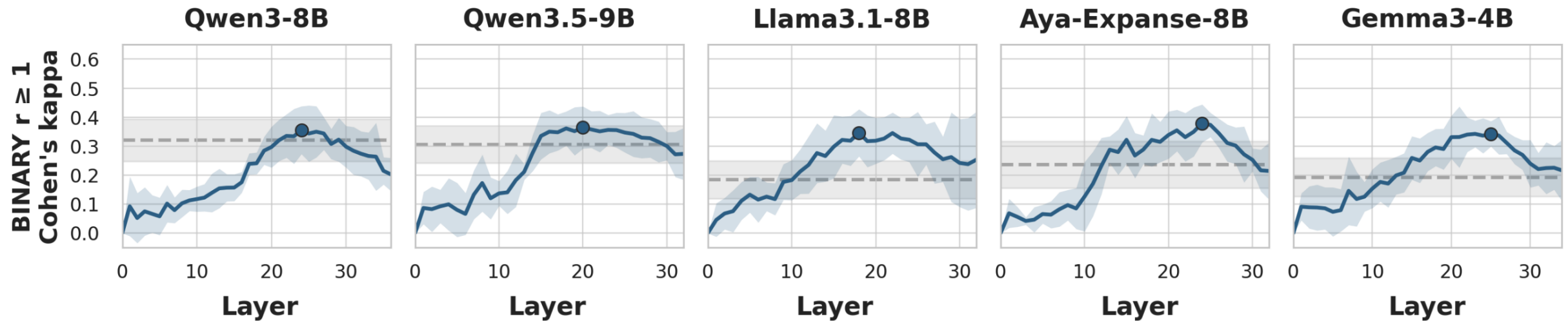


Figure 1: **Layer-wise performance of difficulty probes across languages for LLaMA-3.1-8B.** Heatmap shows, for each test language (rows) and transformer layer (columns), the average Spearman correlation between probe predictions and ground-truth difficulty, where each value is averaged over probes trained on all other languages. Performance peaks in the middle layers, indicating that difficulty representations are most consistently aligned across languages at intermediate depths of the network.

Stefano Civelli, Pietro Bernardelle, Nicolò Brunello, and Gianluca Demartini. **A Shared Geometry of Difficulty in Multilingual Language Models.** In: The 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026). San Diego, California, July 2026.

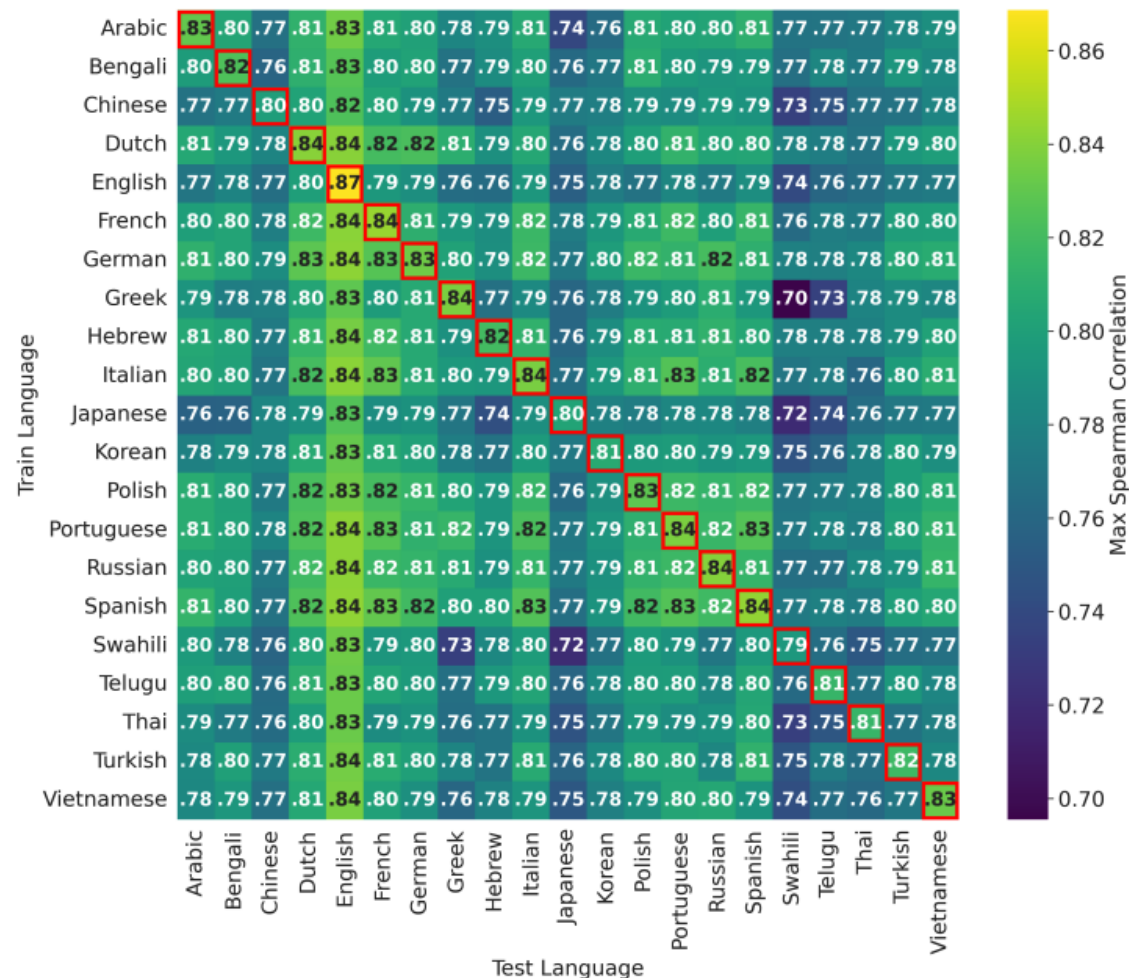
The same is true for LLMs judging relevance



Pietro Bernardelle, Samaneh Mohtadi, Stefano Civelli, Joel Mackenzie, and Gianluca Demartini. **LLMs Encode Relevance as a Layer-Wise Cross-Lingual Signal**. In: The 35th International ACM Conference on Knowledge and Information Management (CIKM 2026). Rome, Italy, November 2026.

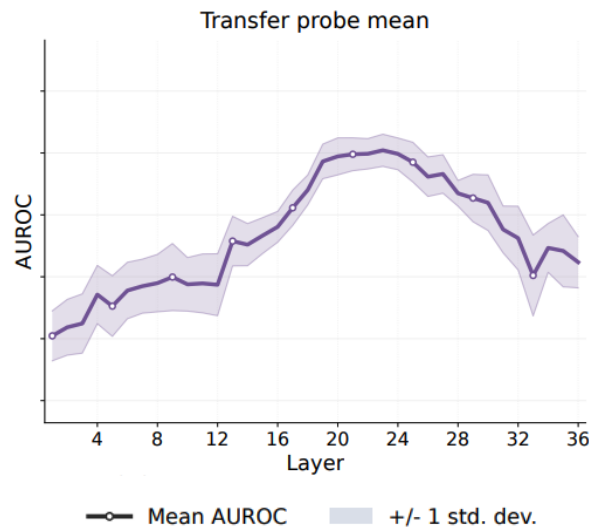
Estimating the difficulty of a prompt

Training and testing difficulty probes across languages works well.



Next steps – work in progress

Use the best internal linear probe to estimate the difficulty of a question / quality of the answer and information about model costs (e.g., price per token) to improve cost/quality results.



(b) Qwen3-8B

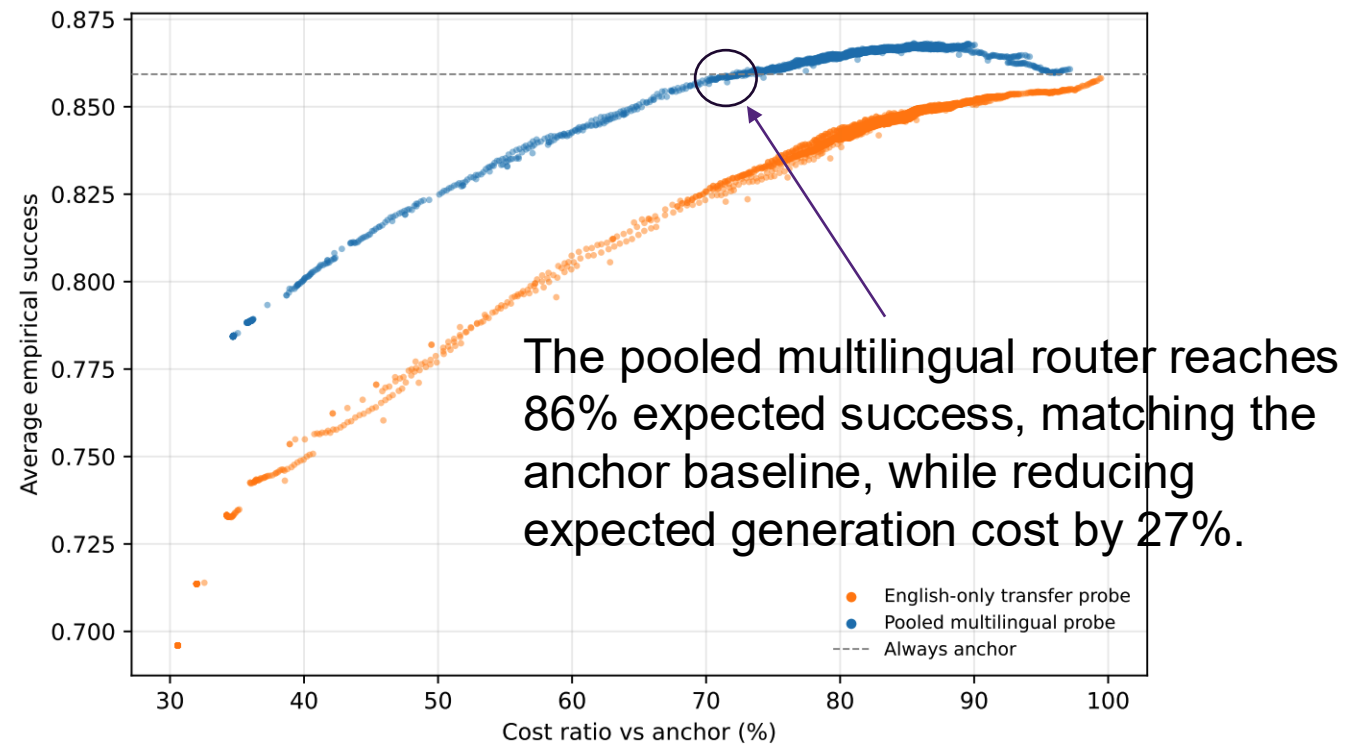


Figure 4.8: Accuracy–cost trade-off for reusable probe-based routing on the test split. Each point corresponds to a routing policy obtained by varying the routing parameters. Cost is reported as a percentage of the expected cost of always using `gpt-oss-20b_high`.

Lessons learned and what to do

- LLMs carry human bias, but are used in decision making
- How do we make humans aware of existing bias?
- How do we build LLMs to better manage bias?
- Increasing transparency, (e.g., track and profile data bias across AI pipelines)
- Select and diversify the sources of data (i.e., diverse human annotators, LLMs)
- Routing to different LLMs
- **Bias management** instead of bias removal



DOI:10.1145/3611641

Gianluca Demartini, Kevin Roitero, and Stefano Mizzaro

Opinion Data Bias Management

*Envisioning a unique approach
toward bias and fairness research.*

Visiting PhD Students Scheme

Visit us in Brisbane, Australia!

2 or 3 months visits for PhD students to work on a joint paper

Funding info and application instructions: <https://d-lab.github.io/visiting/>



Since 2023, we hosted 15+ PhD students based in 8 countries (CH, NL, DE, DK, NO, BE, CN, IT); Example outputs:

Martin Paul Wessel, Timo Spinde, Jürgen Pfeffer, and Gianluca Demartini. **Definitional Sensitivity in Media Bias Detection: A Multi-Definition Dataset and Benchmark.** In: Findings of the 2026 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2026**). Budapest, Hungary, October 2026.

Mads Skipanes, Tollef Emil Jørgensen, Kyle Porter, Gianluca Demartini, and Sule Yildirim Yayilgan. **Enhancing Criminal Investigation Analysis with Summarization and Memory-based Retrieval-Augmented Generation: A Comprehensive Evaluation of Real Case Data.** In: The 31st International Conference on Computational Linguistics (**COLING 2025**).

