



Università degli Studi di Padova

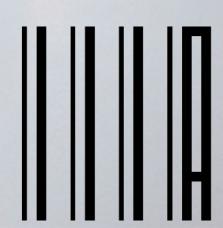


Growing up in Cranfield, Maturing in Generative IR

Nicola Ferro









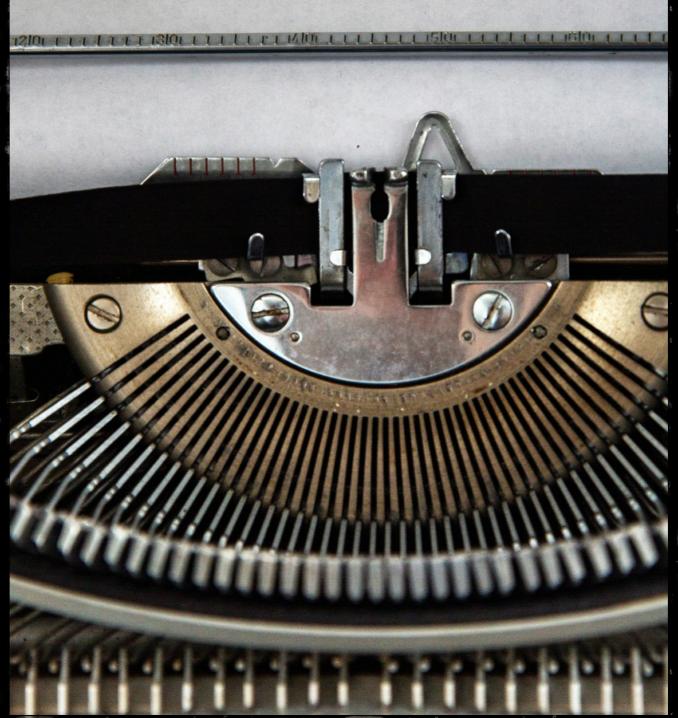
12 -4.4





DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!

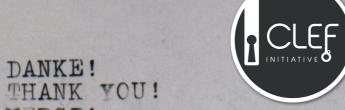
.



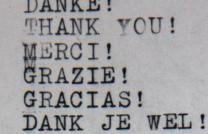


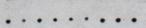




























The Beginning



- Cranfield Paradigm by Cyril W. Cleverdon
- Defines the use of experimental collections
 - documents (corpora)
 - **topics**, which are a surrogate for information needs
 - relevance judgments (binary or graded) also called relevance assessment or ground-truth (or qrels)



Cyril W. Cleverdon

Ensures comparability and repeatability

Cleverdon, C. W. (1962). Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.

Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P. edit

Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, Readings in Information Retrieval, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

of the experiments



The Beginning



- Cranfield Paradigm by Cyril W. Cleverdon
- Defines the use of experimental collections
 - documents (corpora)
 - **topics**, which are a surrogate for information needs
 - relevance judgments (binary or graded) also called relevance assessment or ground-truth (or qrels)



Cyril W. Cleverdon

Ensures comparability and repeatability

of the experiments

Cleverdon, C. W. (1962). Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK. Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, Readings in Information Retrieval, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.



The "Ideal Test Collection"



Corpora → (not historical) corpora are typically OK

- < 500 documents, no real value</p>
- 1-2,000 documents, minimally acceptable
- > 10,000 documents, actually needed

Topics → typical size is still 50 topics

- < 75 topics, no real value</p>
- 250 topics, minimally acceptable
- > 1,000 topics, actually needed

Relevance Judgements → binary is still most common option, diversity only recently

- multi-graded (highly and fairly relevant)
- types (novel, stimulating, ...)
- need for pooling (still open research issue)





Karen Spärck Jones



C. J. "Keith" van Rijsbergen



The "Ideal Test Collection"



Corpora → (not historical) corpora are typically OK

- < 500 documents, no real value</p>
- 1-2,000 documents, minimally acceptable
- > 10,000 documents, actually needed

Topics → typical size is still 50 topics

- < 75 topics, no real value</p>
- 250 topics, minimally acceptable
- > 1,000 topics, actually needed

Relevance Judgements → binary is still most common option, diversity only recently

- multi-graded (highly and fairly relevant)
- types (novel, stimulating, ...)
- need for pooling (still open research issue)



Spärck Jones, K. and van Rijsbergen, C. J. (1975). Report on the need for and provision of an 'ideal' information retrieval test collection. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.



Karen Spärck Jones



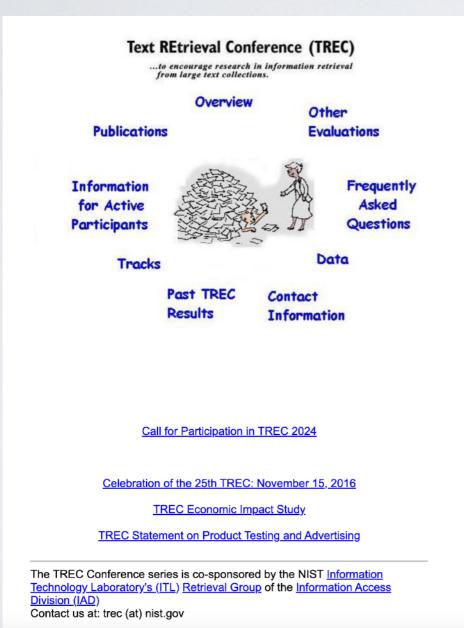
C. J. "Keith" van Rijsbergen



Large-scale Evaluation Initiatives: TREC



- TREC (Text REtrieval Conference), USA, since 1992
 - https://trec.nist.gov/





Ian Soboroff



Donna Harman

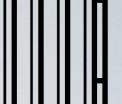


Ellen M. Voorhees

Harman, D. K. and Voorhees, E. M., editors (2005). TREC. Experiment and Evaluation in Information Retrieval. MIT Press, USA.

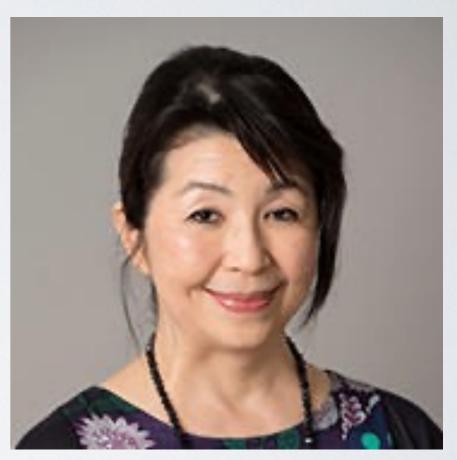


Large-scale Evaluation Initiatives: NTCIR



- NTCIR (NII Testbeds and Community for Information access Research), Japan, since 1999
 - http://research.nii.ac.jp/ntcir/index-en.html





Noriko Kando

Sakai, T., Oard, D. W., and Kando, N., editors (2021). *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany.

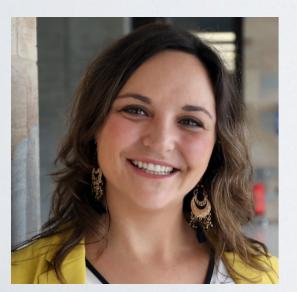


Large-scale Evaluation Initiatives: CLEF

- CLEF (Conference and Labs of the Evaluation Forum),
 Europe, since 2000
 - https://www.clef-initiative.eu/



Alberto Barrón-Cedeño





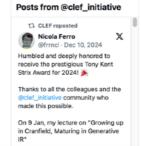
UNLOCKING INFORMATION ACCESS

The **CLEF Initiative** (Conference and Labs of the Evaluation Forum) is structured in two main parts:

- a series of Evaluation Labs, i.e. laboratories to conduct evaluation of information access systems and workshops to discuss and pilot innovative evaluation activities;
- 2. a peer-reviewed Conference on a broad range of issues, including
- investigation continuing the activities of the Evaluation Labs;

The CLEF Initiative promotes research, innovation, and development of information occess systems with an emphasis on multilingual and multimodal information with various levels of structure. CLEF promotes research and development by providing an infrastructure for:

- multilingual and multimodal system testing, tuning and evaluation:
- investigation of the use of unstructured, semistructured, highly-structured, and semantically enriched data in information access:
- creation of reusable test collections for benchmarking;
- exploration of new evaluation methodologies are innovative ways of using experimental data:





Carol Ann Peters



Nicola Ferro

Alba García Seco de Herrera

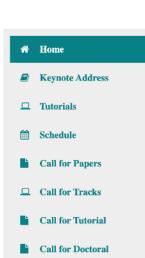
Ferro, N. and Peters, C., editors (2019). *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany.



Large-scale Evaluation Initiatives: FIRE

- FIRE (Forum for Information Retrieval Evaluation),
 India, since 2008
 - https://fire.irsi.org.in/





Registration

Welcome

The 16th meeting of Forum for Information Retrieval Evaluation will be held at DA-IICT, Gandhinagar, India. FIRE was started in 2008 with the aim of building a South Asian counterpart for TREC, CLEF, and NTCIR. It has since evolved continuously to support and encourage research within the information retrieval community and has acted as a unique platform to foster interaction between the researchers of academia and industry. FIRE has continuously adapted to meet the new challenges in multilingual information access and frameworks for large-scale evaluation of information retrieval methods, primarily focusing on text.

In its 16th year, FIRE will continue with the peer-reviewed conference track, shared tasks, industry track, and doctoral consortium. The peer-reviewed conference track aims to encourage novel research among early-stage researchers. The shared tasks aim to bolster research in IR and NLP among the larger research community. The industry track aims to provide the audience with a view of the application of IR research in the industry. The doctoral consortium aims to provide a platform for PhD scholars working in this field and support them through mentorship to improve their research.

Note: Non indian nationals would require visas





SPONSORS



ACM SIGIR

PUBLICATIONS

To be announced soon.



Mandar Mitra



Prasenjit Majumder



Large-scale Evaluation Initiatives: MediaEval



- MediaEval (Benchmarking Initiative for Multimedia Evaluation),
 - Europe, since 2010
 - https://multimediaeval.github.io/



MediaEval 2025 Call for Task Proposals

September 24, 2024

The Multimedia Evaluation Benchmark, MediaEval, offers challenges in artificial intelligence related to data that includes multiple modalities (e.g., audio, visual, textual, and/or contextual). The goal of MediaEval is to develop and evaluate new algorithms and technologies for analyzing, exploring and accessing information in multimedia data. MediaEval pursues a "Quest for Insight": we push beyond improving evaluation scores to achieving deeper understanding about the challenges, including data and the strengths and weaknesses of particular types of approaches. Our larger aim is to promote reproducible research that makes multimedia a positive force for society. MediaEval is now calling for proposals for tasks to run in the 2025 benchmarking season.

- Call for Task proposals (first deadline): Wed. 11 December
- Call for Task proposals (final deadline): Wed. 22 January

The proposal should describe the motivation of the task, including a description of the use scenario in which the results of the tasks would be used (e.g., application that serves users). It should provide a definition of the specific problem that task participants are required to solve. Also, it should include information on the data (including source and licensing), and on how the solutions developed by task participants will be evaluated (the metric and description of how the metric is related to the use scenario). We ask you to think carefully about specific research questions that are related to the challenge, and mention these in the proposal. These research questions will guide participants in pursuing the "Quest for Insight", i.e., going beyond thinking only about evaluation scores. Finally, the proposal must also include a statement of how the task is related to Media Final (i.e., the human or social component) and how it extends the state of the art.



Gareth Jones



Martha Larson



Achievements: A Science of Ms



Methodologies

Measures

Models

Major Systems

Massive Collections



How Valuable is Evaluation?



- The TREC 2010 Economic Impact study estimated in about 30 M\$ the overall investment in TREC by NIST
 - probably much much more if we had a means to estimate also the investment by participants in TREC
- They are the pillars for all the subsequent scientific research and technology development
 - TREC estimated the return on investment in the range of 3\$-5\$ for each invested dollar

Rowe, B. R., Wood, D. W., Link, A. L., and Simoni, D. A. (2010). *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA. http://trec.nist.gov/pubs/2010.economic.impact.pdf.



Courtesy of Paolo Rosso at CLEF 2013



CLIR: The Grand Challenge



AAAI 1997 Spring Symposium: Fully multilingual and multimodal information retrieval systems

- capable of processing a query in any medium and any language
- finding relevant information from a multilingual multimedia collection containing documents in any language and form
- and presenting it in the style most likely to be useful to the user



Doug Oard



David Hull

Oard, D. W. and Hull, D. A. (1997). AAAI 1997 Spring Symposium Series Reports. Cross-Language Text and Speech Retrieval. AI Magazine, 18(3):82.

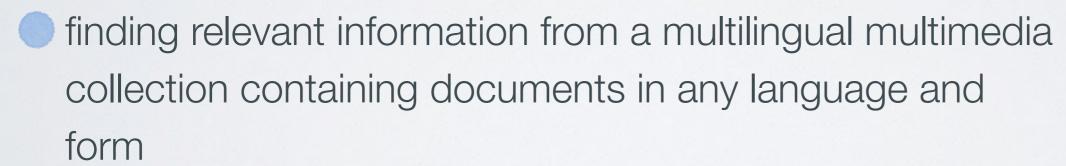


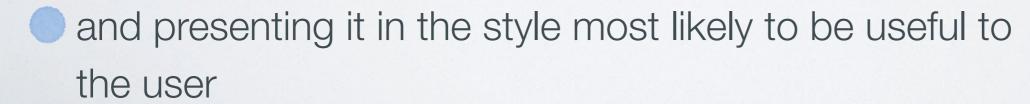
CLIR: The Grand Challenge



AAAI 1997 Spring Symposium: Fully multilingual and multimodal information retrieval systems









Doug Oard



David Hull

Oard, D. W. and Hull, D. A. (1997). AAAI 1997 Spring Symposium Series Reports. Cross-Language Text and Speech Retrieval. AI Magazine, 18(3):82.



Classic CLEF: Years+ of History



- 1997 First CLIR system evaluation campaigns in US and Japan:
 TREC and NTCIR
 - CLEF actually began life in 1997 as a track for Cross Language Information Retrieval (CLIR) within TREC. Mainly, English centered tasks (EN -> X, X -> EN).

- 2000-2009 CLIR evaluation in Europe: CLEF (Cross-Language Evaluation Forum)
 - Fully multilingual, multimodal information retrieval systems capable of processing a query in any medium and any language finding relevant information from a multilingual multimedia collection containing documents in any language and form, and presenting it in the style most likely to be useful to the user



Classic CLEF: Years+ of History



age Information

'EN -> X, X ->

- US and Japan: 1997 - First CLIR system evaluation campaign TREC and NTCIR
 - CLEF actually began Retrieval (CLIP) EN).
 - Multingual entertains and the second of the Nah Guages

 Nah Guages

 Medir 2000-2009 nguage Evaluation F
 - Fully multiling cems capable of processing a y language finding relevant information from nal multimedia collection containing documents in and presenting it in the style most likely to be useful any language ar to the user



Carol: The True Name Behind CLEF







Where To Go Next? The CLEF Initiative



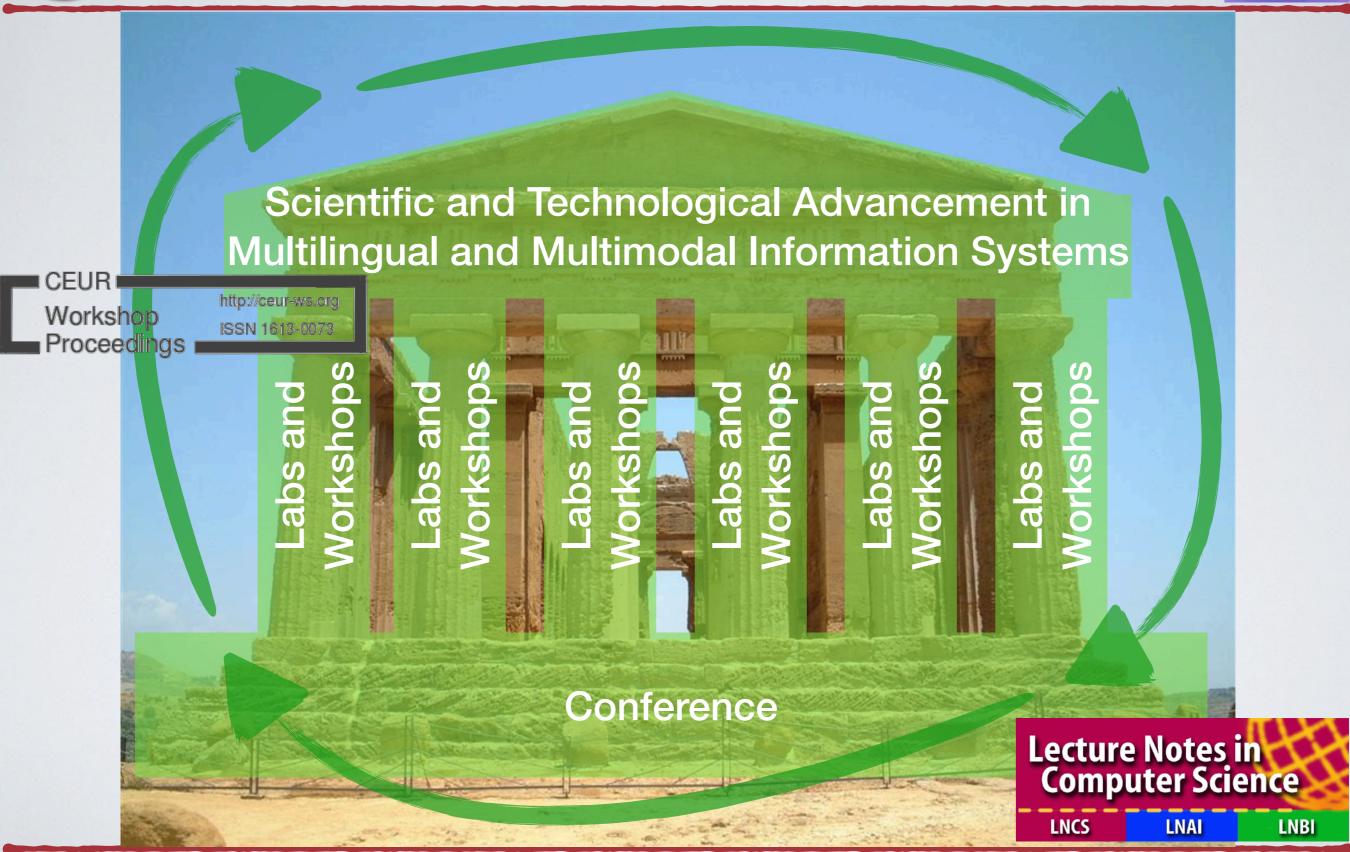
The direction depended on the Fellowship

The CLEF Initiative is a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure.



Approach (since 2010)







CLEF and ECIR (since 2019)







CLEF Resultless Review Process (since 2023)



Dagstuhl Seminar 23031, January 2023 Frontiers of Information Access Experimentation for Research and Education



Bauer, C., Carterette, B. A., Ferro, N., Fuhr, N., and Faggioli, G., editors (2023). Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education, Dagstuhl Reports, Volume 13, Number 1. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Germany.

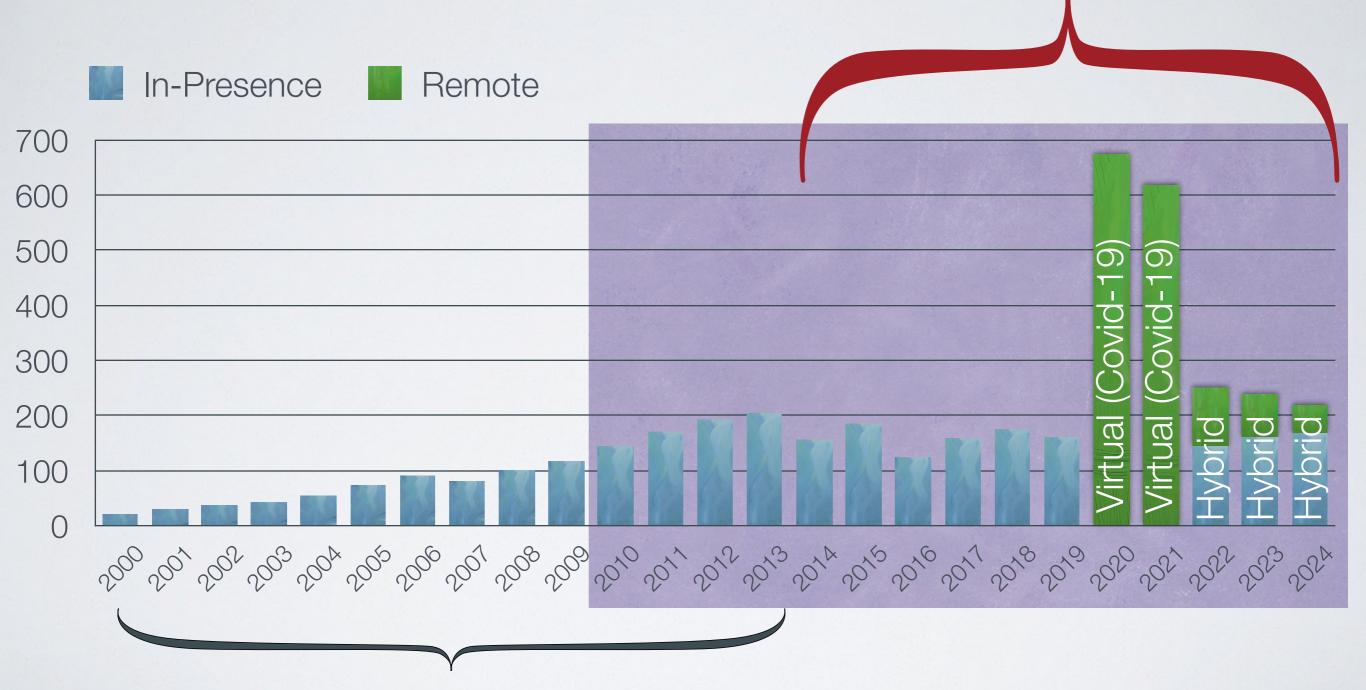




Participation: Attendees



100% Voluntary Effort Based

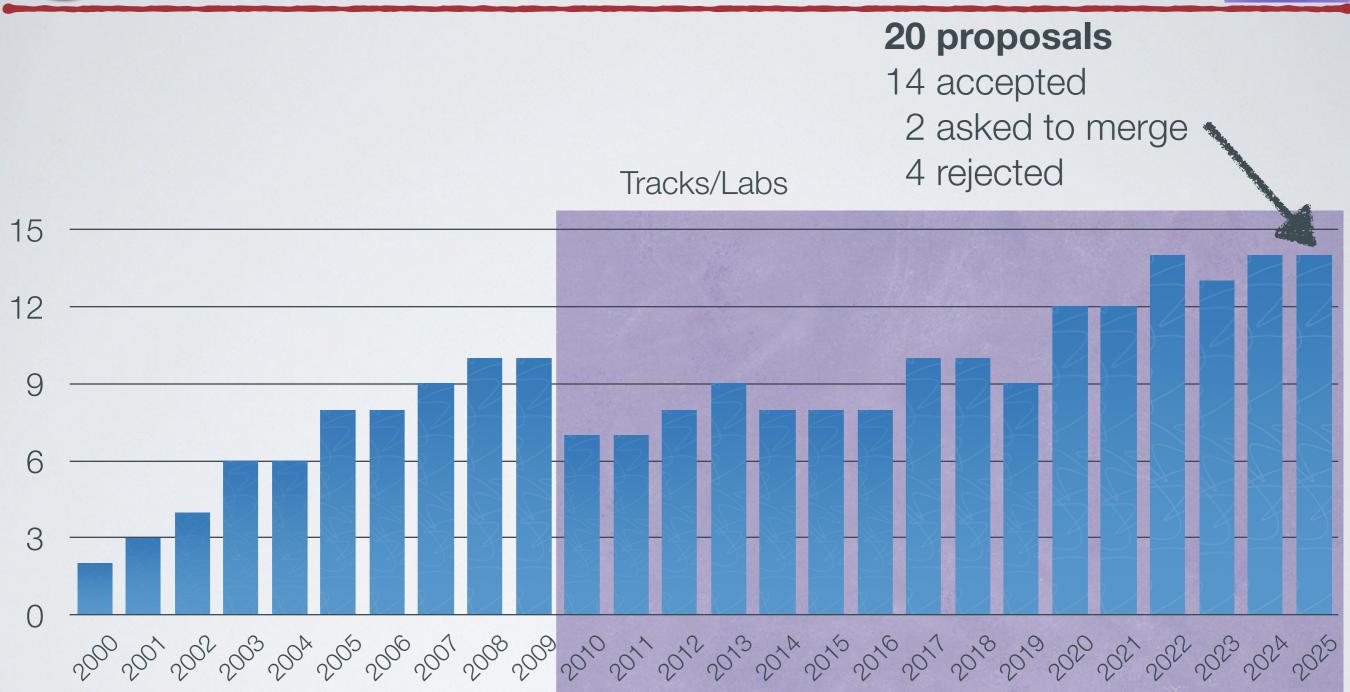


Mainly voluntary effort + project funding



Labs



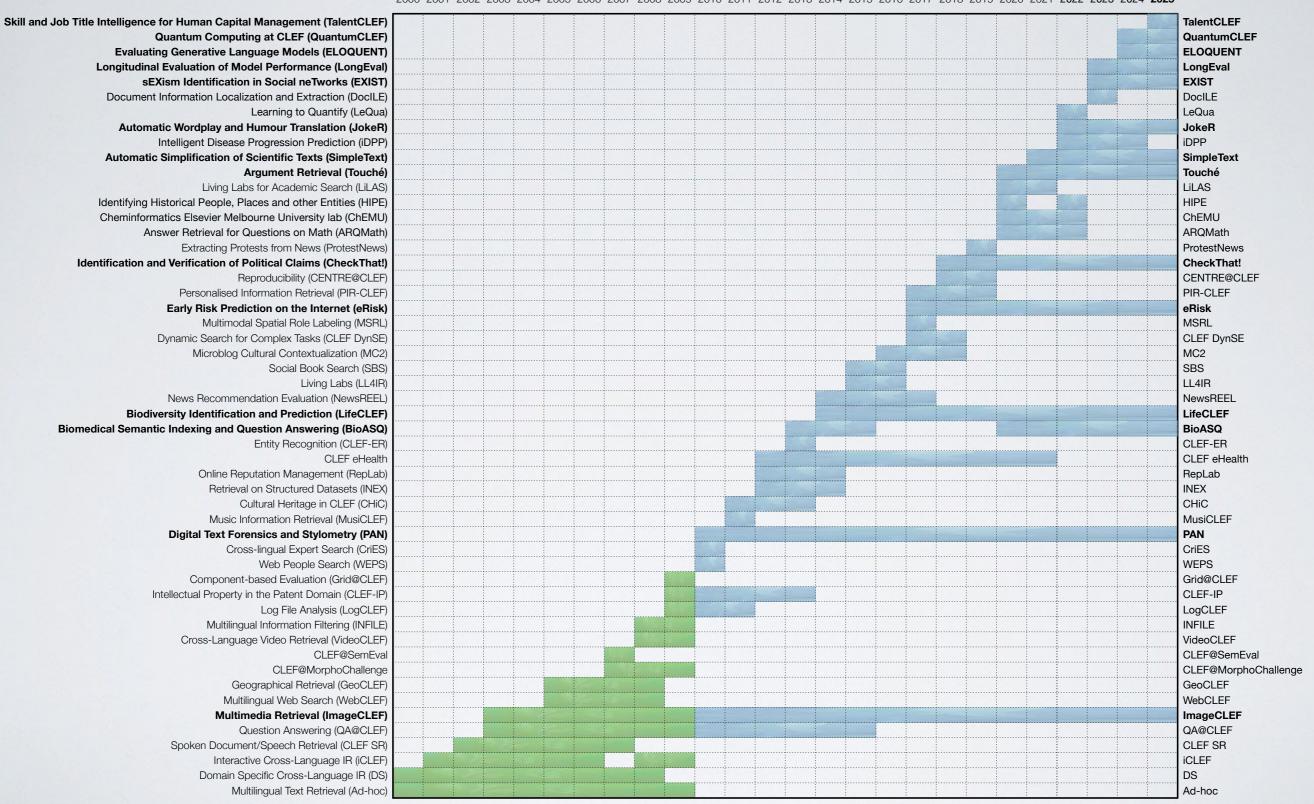




CLEF Labs over Time



2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 **2022 2023 2024 2025**

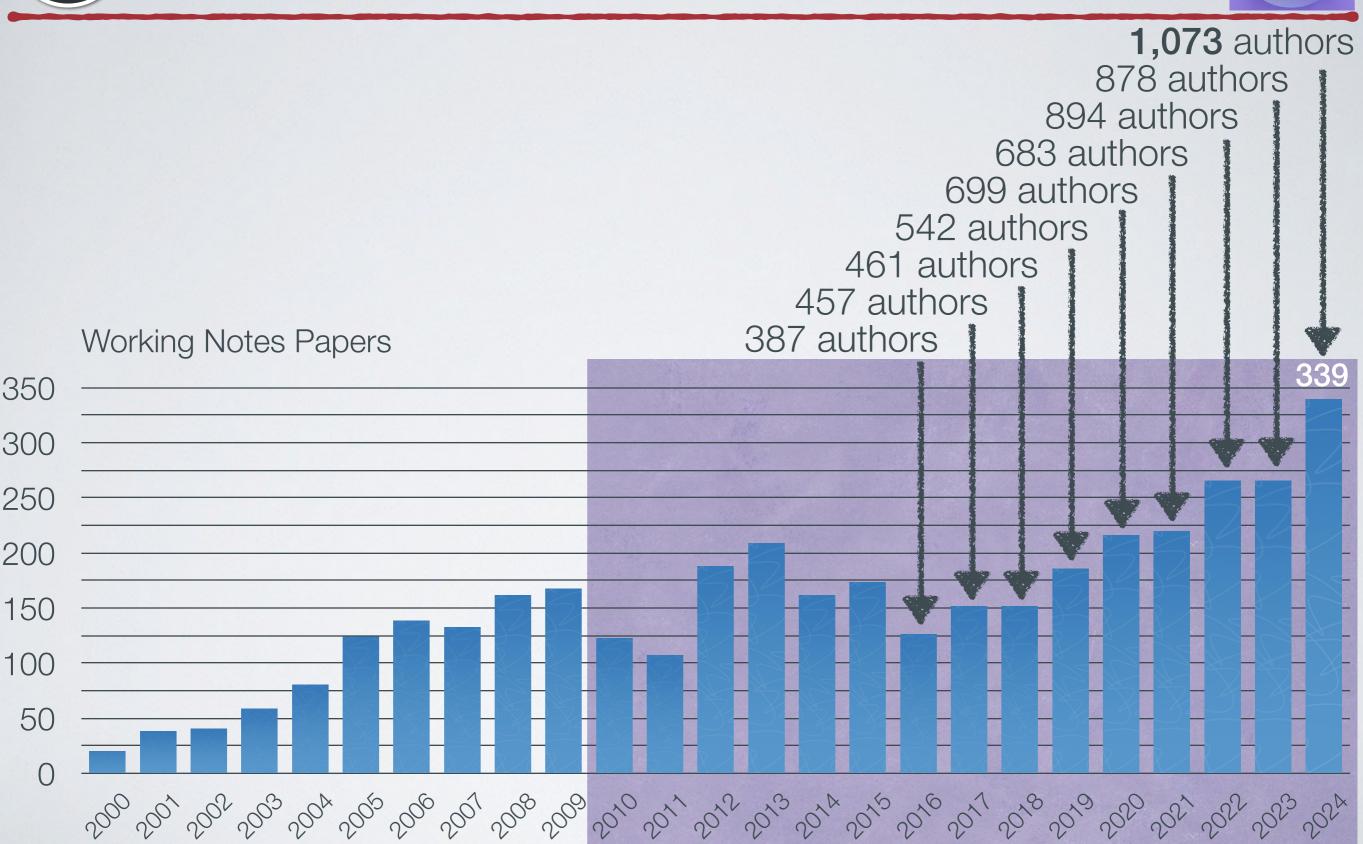


2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 **2022 2023 2024 2025**



Working Notes



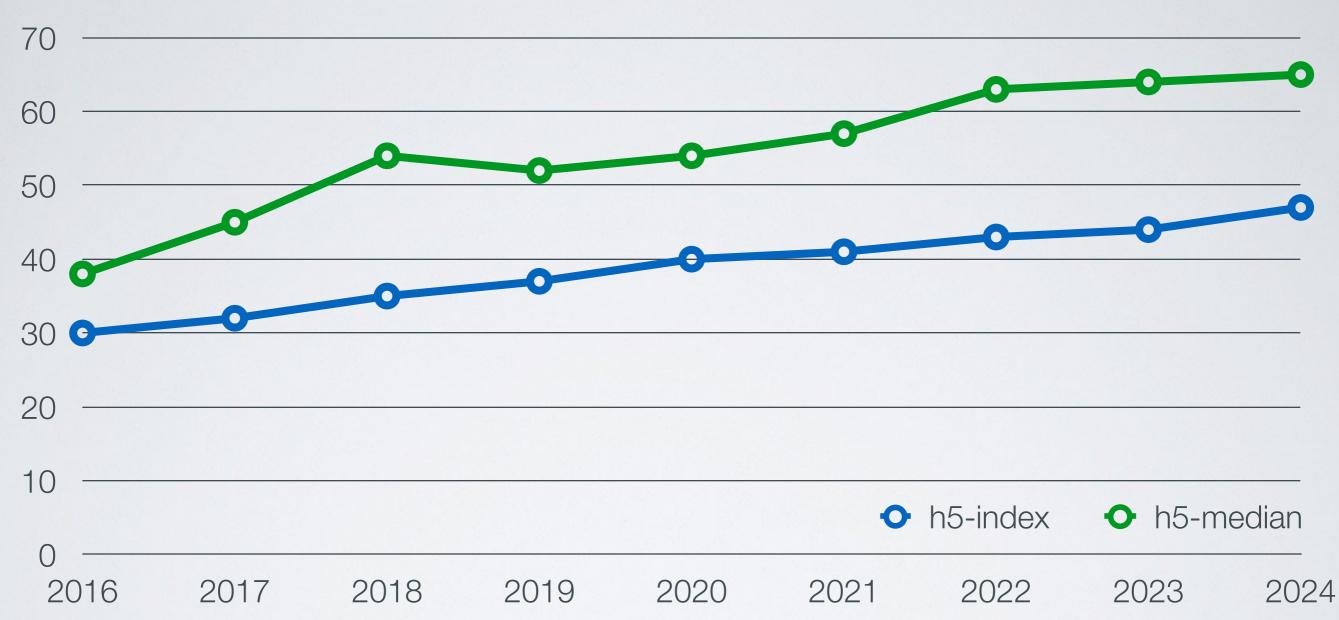




Publication "Universe"



Google Scholar Metrics for "Cross-Language Evaluation Forum"



- Google Scholar for "CLEF evaluation"
 - 96,300 hits



Publication

Publication "Universe" (2024)

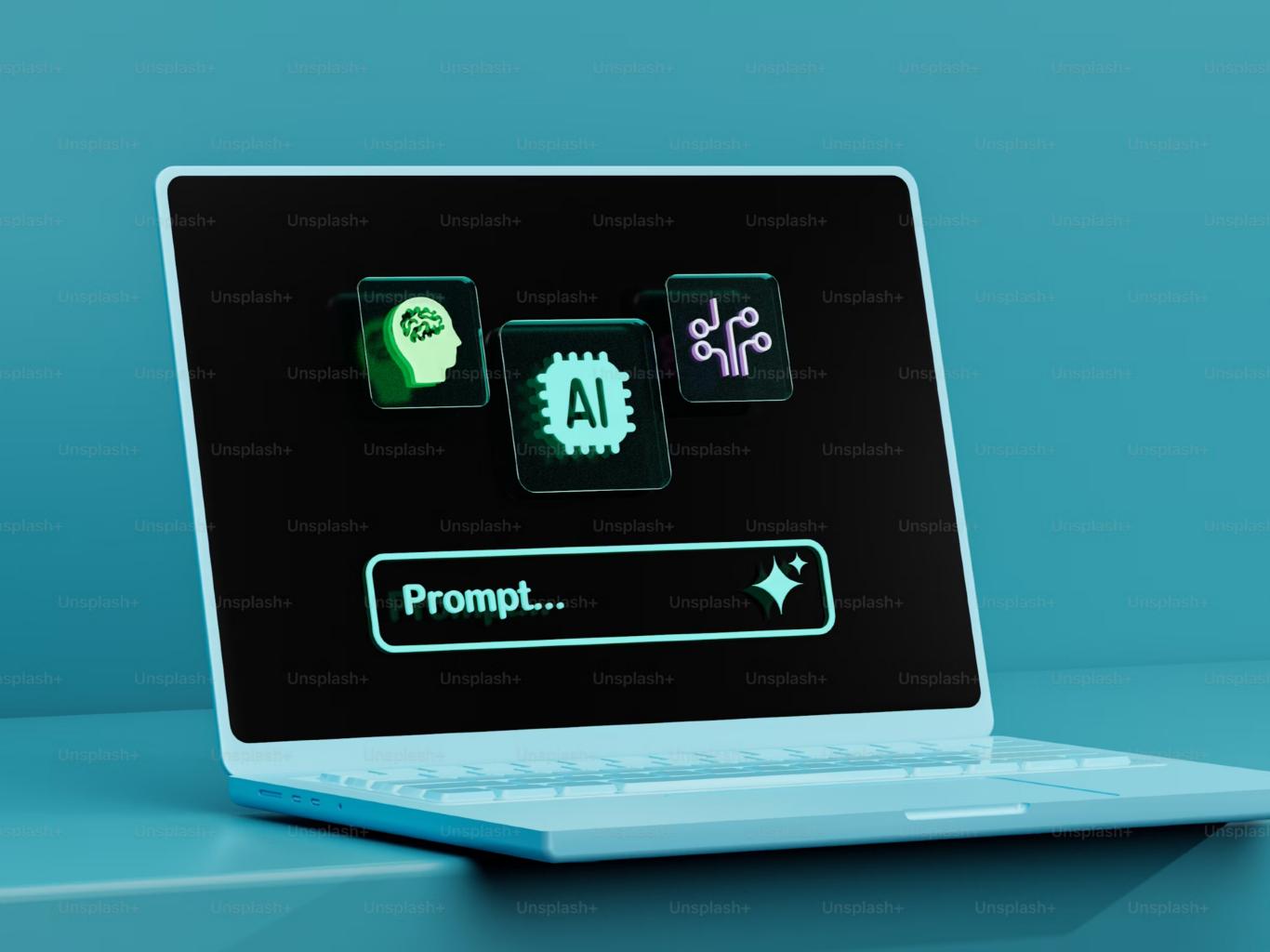


"AIRS" OR "WWW" OR "information retrieval" OR "Information and Knowledg Q

Top 20 publications matching "AIRS" OR "WWW" OR "information retrieval...

		Publication	h5-index	<u>h5-median</u>
	1.	International World Wide Web Conferences (WWW)	112	169
	2.	IEEE Transactions on Knowledge and Data Engineering	<u>107</u>	185
	3.	ACM SIGIR Conference on Research and Development in Information Retrieval	<u>103</u>	149
	4.	Information Processing & Management	<u>96</u>	157
	5.	ACM International Conference on Information and Knowledge Management	<u>91</u>	133
	6.	ACM International Conference on Web Search and Data Mining	<u>77</u>	130
<u>h5-median</u>	7.	ACM Conference on Recommender Systems	<u>53</u>	81
169	8.	Journal of the Association for Information Science and Technology	<u>49</u>	69
185	9.	ACM Transactions on Information Systems (TOIS)	<u>48</u>	98
149	10.	Workshop of Cross-Language Evaluation Forum	<u>47</u>	65
157	11.	International Society for Music Information Retrieval Conference	<u>43</u>	69
133	12.	European Conference on Advances in Information Retrieval	<u>42</u>	60
191	13.	VINE Journal of Information and Knowledge Management Systems	<u>38</u>	61
122	14.	Forum for Information Retrieval Evaluation	<u>30</u>	39
130	15.	ACM SIGIR International Conference on Theory of Information Retrieval	<u>24</u>	42
109	16.	ACM Transactions on the Web (TWEB)	<u>24</u>	39
94	17.	International Journal of Multimedia Information Retrieval	<u>24</u>	36
88	18.	International ACM/IEEE Joint Conference on Digital Libraries	<u>23</u>	33
79	19.	Romanian Journal of Information Science and Technology	<u>19</u>	32
81	20.	International Journal on Digital Libraries	<u>19</u>	30

1.	International World Wide Web Conferences (WWW)	<u>112</u>	169
2.	IEEE Transactions on Knowledge and Data Engineering	<u>107</u>	185
3.	ACM SIGIR Conference on Research and Development in Information Retrieval	<u>103</u>	149
4.	Information Processing & Management	<u>96</u>	157
5.	ACM International Conference on Information and Knowledge Management	<u>91</u>	133
6.	Journal of Big Data	<u>79</u>	191
7.	International Conference on Very Large Databases	<u>79</u>	122
8.	ACM International Conference on Web Search and Data Mining	<u>77</u>	130
9.	ACM SIGMOD International Conference on Management of Data	<u>73</u>	109
10.	International Conference on Data Engineering	<u>69</u>	94
11.	International Conference on Web and Social Media (ICWSM)	<u>56</u>	88
12.	IEEE International Conference on Big Data	<u>54</u>	79
13.	ACM Conference on Recommender Systems	<u>53</u>	81
14.	Information Systems	<u>49</u>	77
15.	World Wide Web	<u>49</u>	71
16.	ACM Transactions on Information Systems (TOIS)	<u>48</u>	98
17.	Knowledge and Information Systems	<u>48</u>	85
18.	ACM Transactions on Intelligent Systems and Technology (TIST)	<u>47</u>	85
19.	Workshop of Cross-Language Evaluation Forum	<u>47</u>	65
20.	ACM Transactions on Internet Technology (TOIT)	<u>44</u>	69





R.I.P. Information Retrieval?







Who is Helping Who?





Retrieval Augmented Generation

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Proc. 34th Annual Conference on Neural Information Processing Systems (NeurlPS 2020)*, pages 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020.



Do We Always Need Generation?





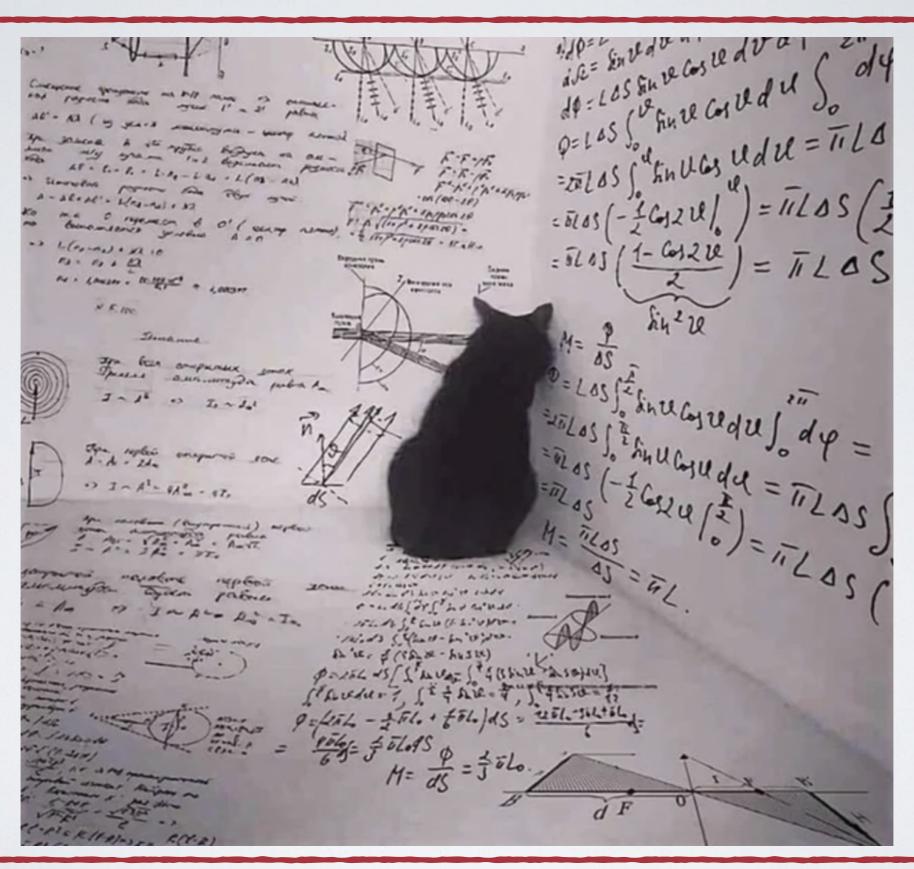
Generation





How to Adapt Cranfield for Evaluating RAG?







Why Not Just Use LLMs for Relevance Assessment?



- Judgment bias toward a particular LLM
- Bias toward user groups
- Resilience against misinformation
- LLM-based LLM training
- Judging vs. predicting
- Truthfulness and hallucinations



A Spectrum of Human-LLM/AI Collaboration



Collaboration perspective: Spectrum of possibilities for collaborative \dagger human— \clubsuit machine task organization to make (relevance) decisions. The \triangle symbol indicates where on the spectrum each possibility falls.

Collaboration Balance		Task Allocation
Human Judgment		
	- 💭	Humans manually decide what is relevant without any kind of AI support.
	-\$	Humans have full control of deciding but are supported by machine-based text highlighting, data clustering, and so forth.
Model in the Loop		
†	- 5	Humans decide based on LLM-generated summaries needed for the decision.
	- \$	Balanced competence partitioning. Humans and LLMs focus on decisions they are good at.
Human in the Loop		
*	- \$ \$	Two (or more) LLMs each generate a decisio and a human selects the best one.
	- 😘	An LLM makes a decision (and an explanatio for it) that a human can accept/reject.
	— (5) • n	LLMs are considered crowdworkers varied by specific characteristics, aggregrated, and controlled by a human.
Fully Automated		
		Fully automatic decision without humans.

- Human judgment. On one extreme, humans make all relevance judgments manually without being influenced by an LLM
- Model in the loop. To make it easier for human assessors to decide on relevance in a consistent manner, an advanced level of automatic support could be provided, e.g. summarizing documents
- Human in the loop. Automated judgments could be produced by an LLM and then verified by humans
- Fully automated. If LLMs were able to reliably judge relevance, they could completely replace humans when judging relevance.

Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., and Wachsmuth, H. (2024). Who determines what is relevant? Humans or Al? Why not both! A Spectrum of Human-Al Collaboration in Assessing Relevance. *Communications of the ACM (CACM)*, 67(4):31–34.



Reproducibility





Different Data/Setup Same Task/Goal

Same/Different Software

Different Group

Reproduce

Same Data/Setup Same Task/Goal Same Software Different Group



Repeat

Same Data/Setup Same Task/Goal Same Software Same Group

Generalize

Different Data/Setup

Different Task/Goal

Same/Different Software

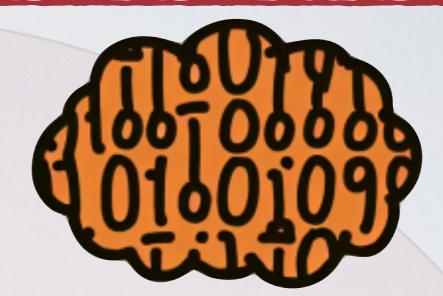
Different Group



Reproducibility





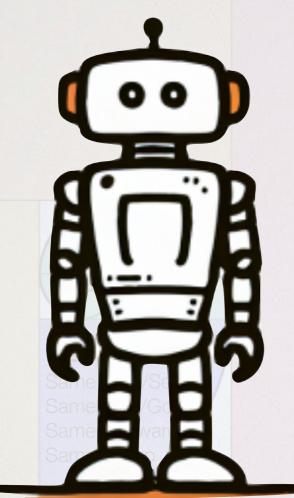


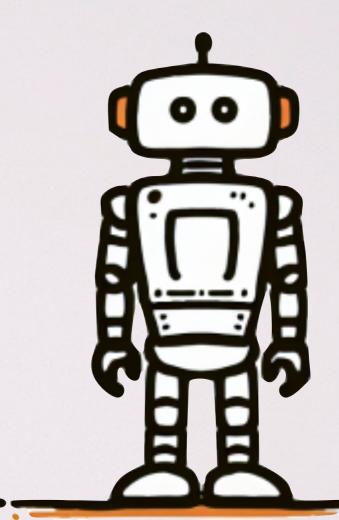
Replicate

Different Data/Setup
Same Task/Goal
Same/Different Software
Different Group

Reproduce

Same Data/Setup Same Task/Goal Same Software Different Group





Generalize

Different Data/Setup

Different Task/Goal

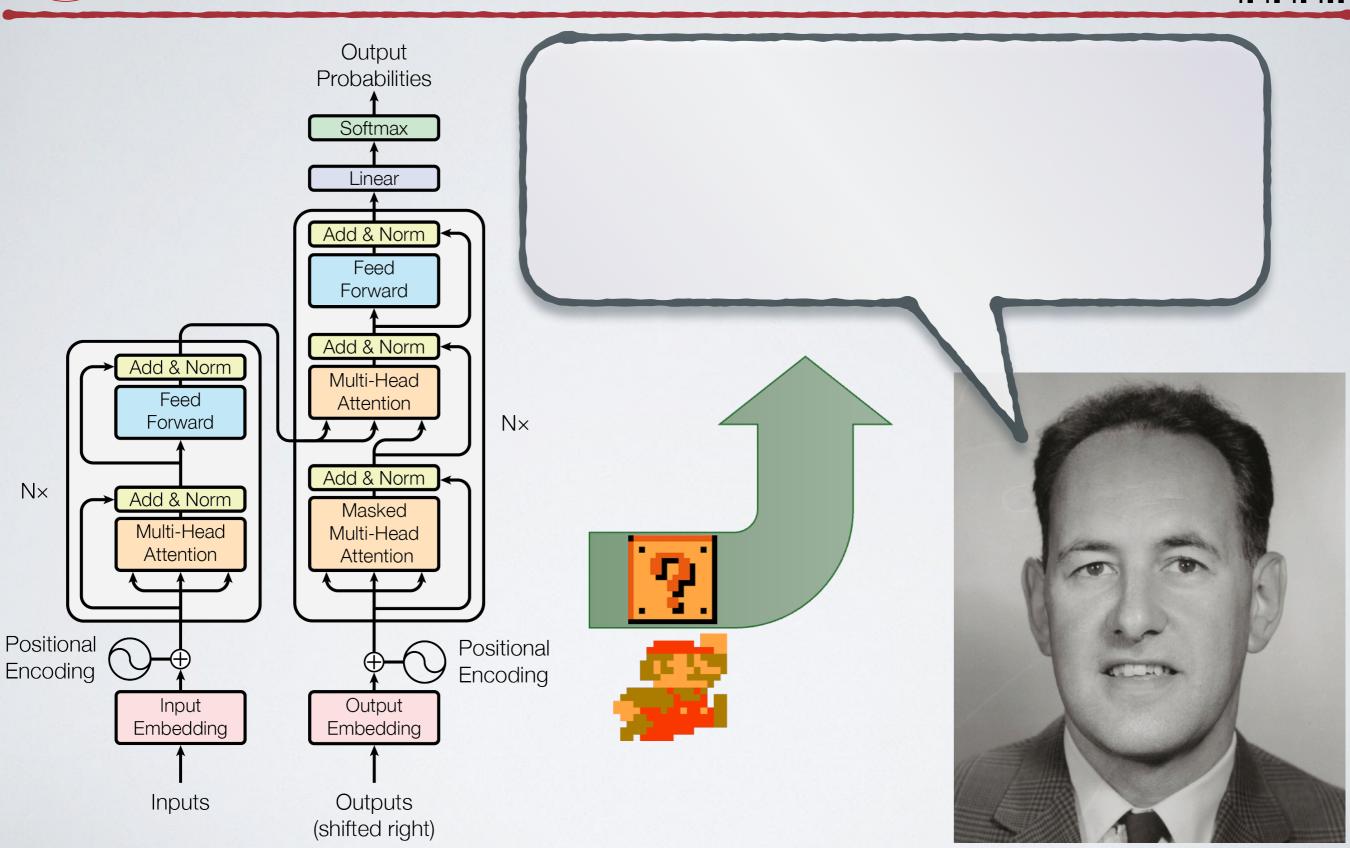
Same/Different Softwar

Different Group



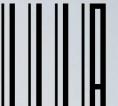
Back to the Roots: What is IR Today?







Back to the Roots: What is IR Today?







Back to the Roots: What is IR Today?







Evaluation Beyond the Surface







Thank You!