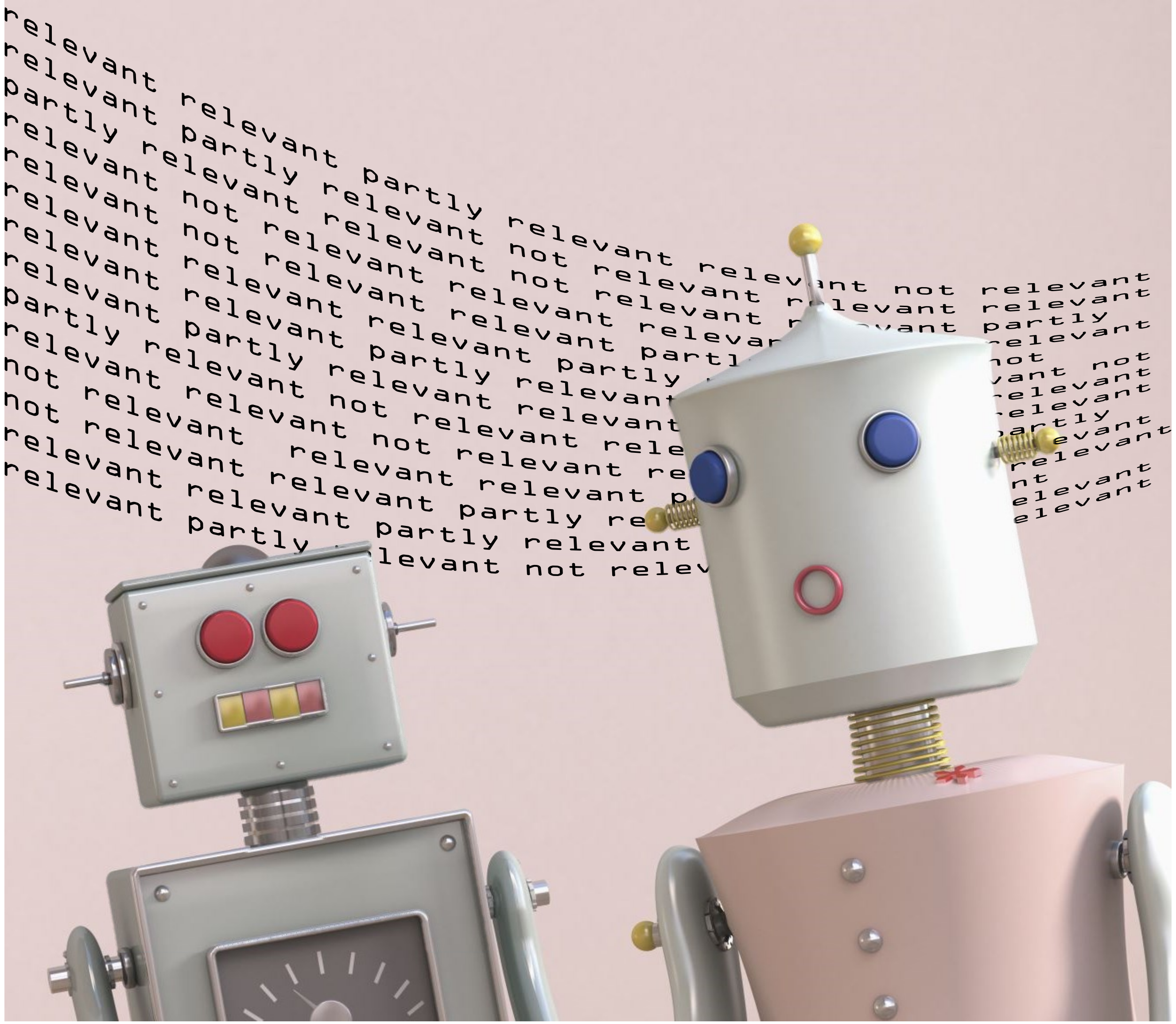
Large language models *for* relevance labelling

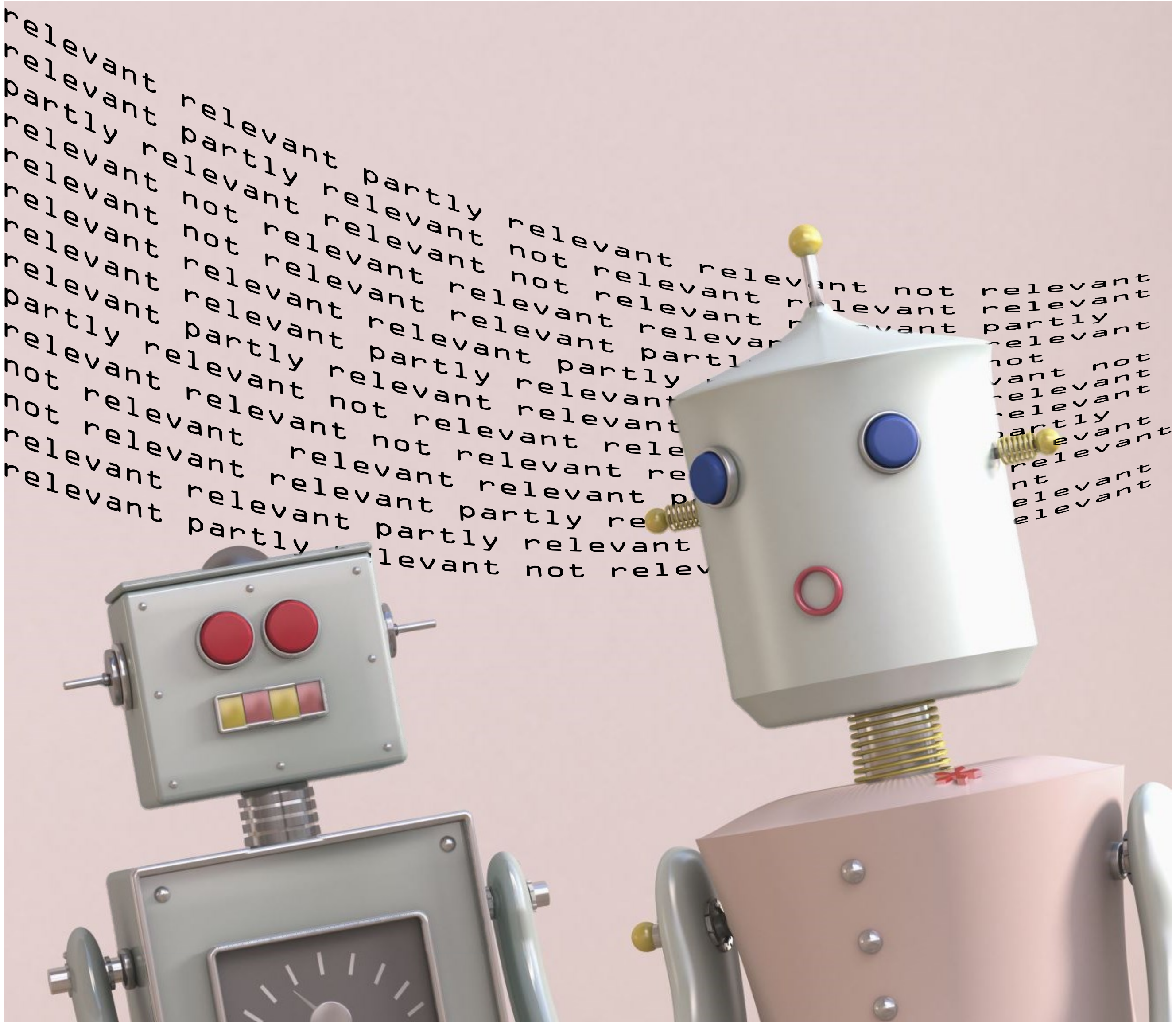


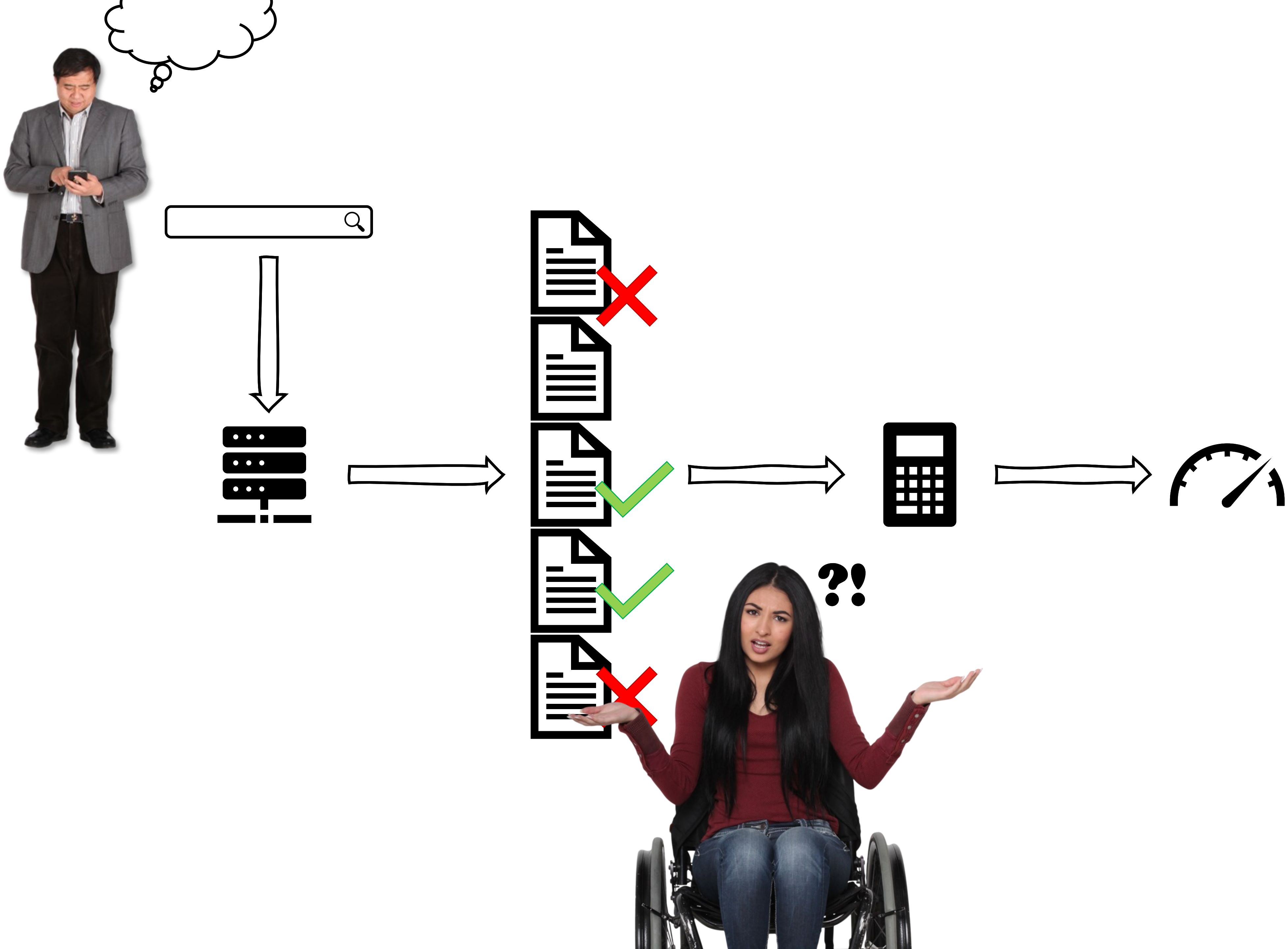
Paul Thomas

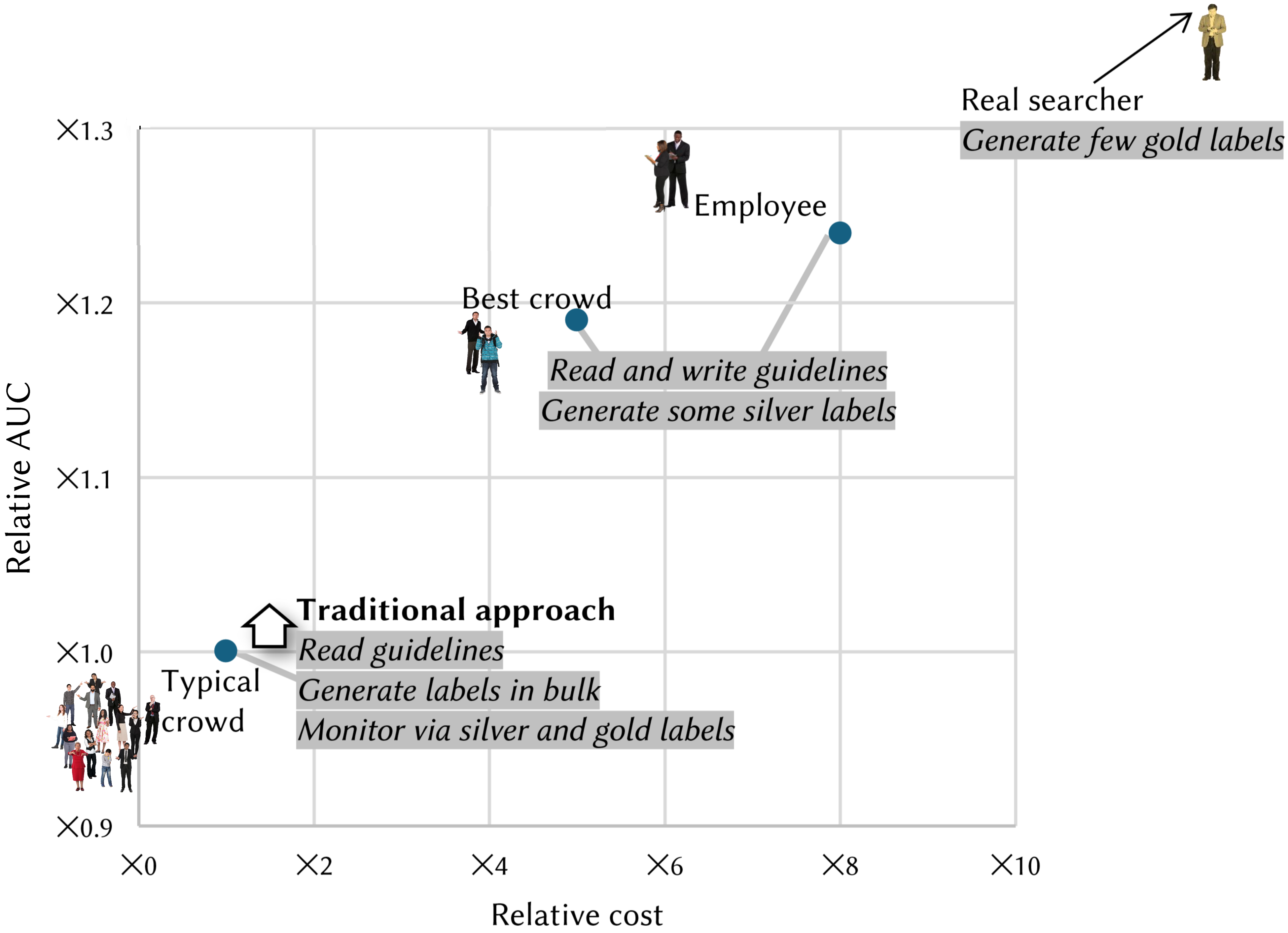
Seth Spielman

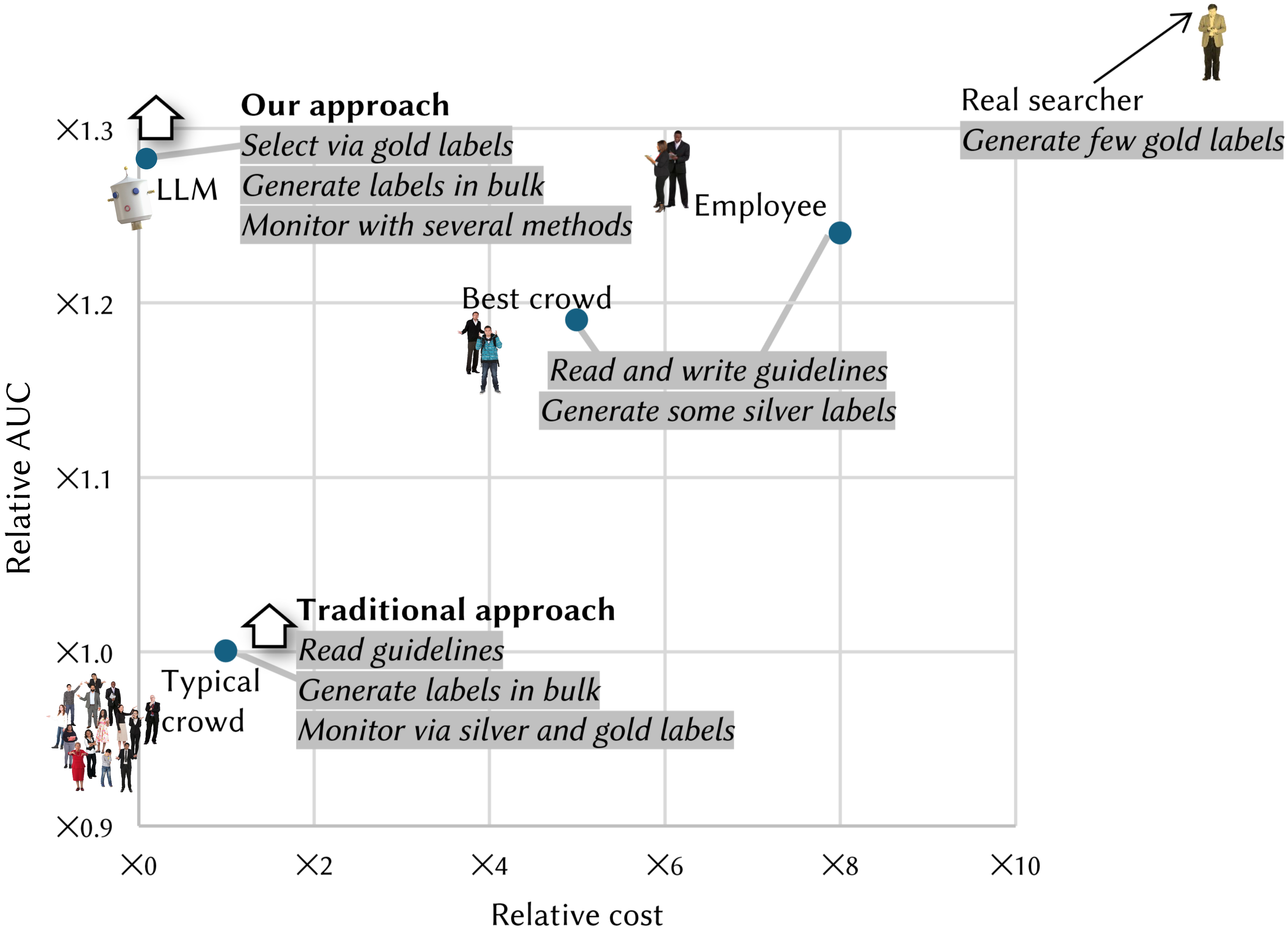
Nick Craswell

Bhaskar Mitra

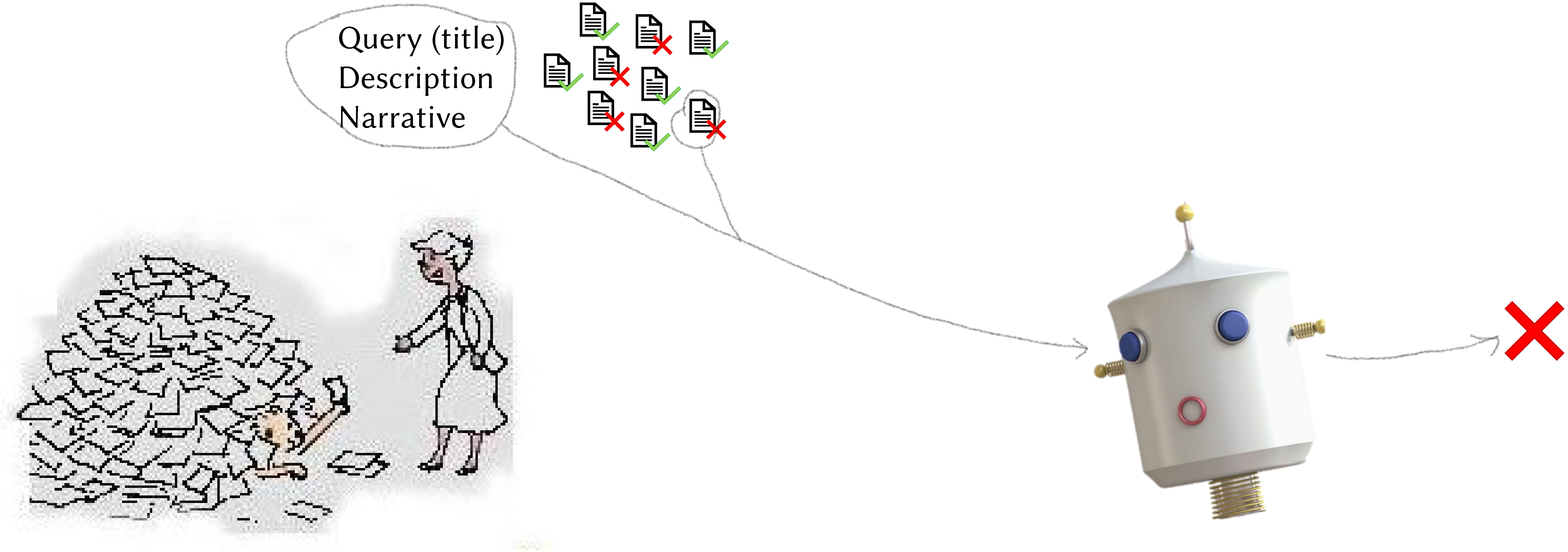





Real searcher
Generate few gold labels
Employee
Best crowd
Read and write guidelines
Generate some silver labels
Traditional approach
Read guidelines
Generate labels in bulk
Monitor via silver and gold labels
Typical crowd
Relative AUC
×1.3
×1.2
×1.1
×1.0
×0.9
×0
×2
×4
×6
×8
×10
Relative cost


Real searcher
Generate few gold labels
Our approach
Select via gold labels
Generate labels in bulk
Monitor with several methods
LLM
Employee
Best crowd
Read and write guidelines
Generate some silver labels
Traditional approach
Read guidelines
Generate labels in bulk
Monitor via silver and gold labels
Typical crowd
×1.3
×1.2
×1.1
×1.0
×0.9
Relative AUC
×0
×2
×4
×6
×8
×10
Relative cost

Some experiments



Query (title)
Description
Narrative

You are a search quality rater evaluating the relevance of web pages. Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

2 = highly relevant, very helpful for this query

1 = relevant, may be partly helpful but might contain other irrelevant content

0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

Query

A person has typed [*query*] into a search engine.

They were looking for: *description narrative*

Result

Consider the following web page. ...

Instructions

Split this problem into steps:

Consider the underlying intent of the search.

Measure how well the content matches a likely intent of the query (M).

Measure how trustworthy the web page is (T).

Query

A person has typed [*query*] into a search engine.

They were looking for: *description narrative*

Result

Consider the following web page. ...

Instructions

Split this problem into steps:

Consider the underlying intent of the search.

Measure how well the content matches a likely intent of the query (M).

Measure how trustworthy the web page is (T).

Consider the aspects above and the relative importance of each, and decide on a final score (O).

We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.

Produce a JSON array of scores without providing any reasoning. Example: [{"M": 2, "T": 1, "O": 1}, {"M": 1...

Results

[{

You are a search quality rater evaluating the relevance of web pages. Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

Role, “R”

2 = highly relevant, very helpful for this query

1 = relevant, may be partly helpful but might contain other irrelevant content

0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

Query

A person has typed [*query*] into a search engine.

They were looking for: *description narrative*

Description, “D”

Narrative, “N”

Result

Consider the following web page. ...

Instructions

Split this problem into steps:

Consider the underlying intent of the search.

Measure how well the content matches a likely intent of the query (M).

Measure how trustworthy the web page is (T).

Query

A person has typed [*query*] into a search engine.

They were looking for: *description narrative*

Description, “D”

Narrative, “N”

Result

Consider the following web page. ...

Instructions

Split this problem into steps:

Consider the underlying intent of the search.

Aspects, “A”

Measure how well the content matches a likely intent of the query (M).

Measure how trustworthy the web page is (T).

Consider the aspects above and the relative importance of each, and decide on a final score (O).

We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.

Produce a JSON array of scores without providing any reasoning. Example: [{"M": 2, "T": 1, "O": 1}, {"M": 1...

Results

Multiple, “M”

[{

Results

Mean absolute error (=1-accuracy, if binary)

Cohen's κ

Area under the ROC curve (AUC, pairwise correctness)

	Document label: MAE	Document label: κ	Document pref: AUC
— — — — —	0.34 ± 0.01	0.38 ± 0.02	0.73 ± 0.01
R — — — —	0.38 ± 0.02	0.32 ± 0.02	0.71 ± 0.01
— D — — —	0.36 ± 0.02	0.35 ± 0.03	0.72 ± 0.01
— — N — —	0.35 ± 0.02	0.37 ± 0.03	0.73 ± 0.01
— — — A —	0.19 ± 0.02	0.60 ± 0.03	0.82 ± 0.02
— — — — M	0.46 ± 0.02	0.22 ± 0.02	0.65 ± 0.01
R D — — —	0.40 ± 0.02	0.30 ± 0.03	0.69 ± 0.01
R — N — —	0.38 ± 0.02	0.33 ± 0.02	0.71 ± 0.01
R — — A —	0.21 ± 0.02	0.56 ± 0.03	0.81 ± 0.02
R — — — M	0.49 ± 0.02	0.20 ± 0.02	0.64 ± 0.01
— D N — —	0.35 ± 0.02	0.37 ± 0.02	0.74 ± 0.01
— D — A —	0.19 ± 0.01	0.59 ± 0.03	0.83 ± 0.01
— D — — M	0.45 ± 0.01	0.24 ± 0.02	0.66 ± 0.01
— — N A —	0.18 ± 0.01	0.62 ± 0.02	0.84 ± 0.01
— — N — M	0.41 ± 0.02	0.29 ± 0.02	0.69 ± 0.01
— — — A M	0.31 ± 0.02	0.42 ± 0.04	0.80 ± 0.02

	Document label: MAE	Document label: κ	Document pref: AUC
R D N — —	0.37 ± 0.02	0.34 ± 0.03	0.72 ± 0.02
R D — A —	0.22 ± 0.01	0.53 ± 0.03	0.82 ± 0.01
R D — — M	0.46 ± 0.02	0.23 ± 0.02	0.66 ± 0.01
R — N A —	0.20 ± 0.01	0.59 ± 0.03	0.83 ± 0.01
R — N — M	0.42 ± 0.02	0.28 ± 0.02	0.69 ± 0.01
R — — A M	0.38 ± 0.02	0.32 ± 0.02	0.78 ± 0.01
— D N A —	0.17 ± 0.01	0.64 ± 0.02	0.85 ± 0.01
— D N — M	0.40 ± 0.02	0.31 ± 0.02	0.70 ± 0.01
— D — A M	0.31 ± 0.01	0.42 ± 0.02	0.80 ± 0.01
— — N A M	0.27 ± 0.02	0.49 ± 0.03	0.82 ± 0.02
R D N A —	0.19 ± 0.01	0.61 ± 0.02	0.84 ± 0.01
R D N — M	0.41 ± 0.01	0.29 ± 0.02	0.69 ± 0.01
R D — A M	0.37 ± 0.02	0.34 ± 0.02	0.80 ± 0.01
R — N A M	0.33 ± 0.01	0.39 ± 0.02	0.80 ± 0.01
— D N A M	0.26 ± 0.01	0.50 ± 0.02	0.82 ± 0.01
R D N A M	0.16 ± 0.02	0.51 ± 0.06	0.77 ± 0.03

Role, R	κ −0.04
Description, D	κ +0.01
Narrative, N	κ +0.06
Aspects, A	κ +0.21
Multiple, M	κ −0.13

Role, R	κ −0.04
Description, D	κ +0.01
Narrative, N	κ +0.06
Aspects, A	κ +0.21
Multiple, M	κ −0.13

R + N	κ −0.01
D + N	κ +0.02
R + A	κ +0.03
D + A	κ +0.02
N + A	κ +0.02
R + M	κ +0.03
D + M	κ +0.04
...etc...	
R + A + M	κ −0.09
R + D + M	κ +0.01

	Hardest query	**Best run**	**Best group**
	Norm. RBO, φ=.9	Norm. RBO, φ=.7	Norm. RBO, φ=.7
P@10	0.40	0.79	0.97
RBP@100, φ=.6	0.42	0.63	0.91
MAP@100	0.48	0.50	0.58
Random	0.04	0.03	0.21

CAUTION!

Binarised labels

One model, few prompts

Other things matter

Observations

Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:
2 = highly relevant, very helpful for this query
1 = relevant, may be partly helpful but might contain other irrelevant content
0 = not relevant, should never be shown for this query
Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

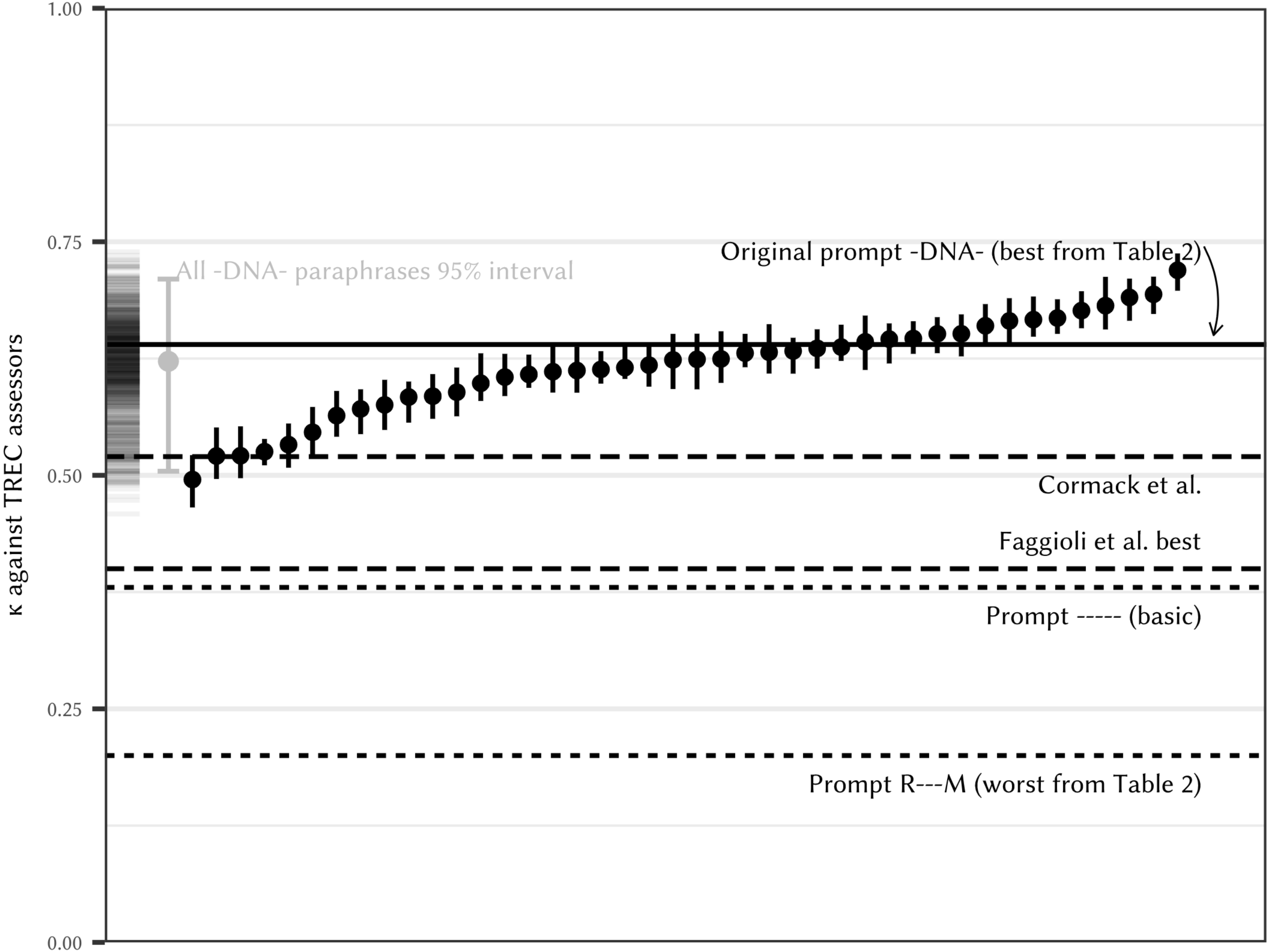
$\kappa = 0.64$

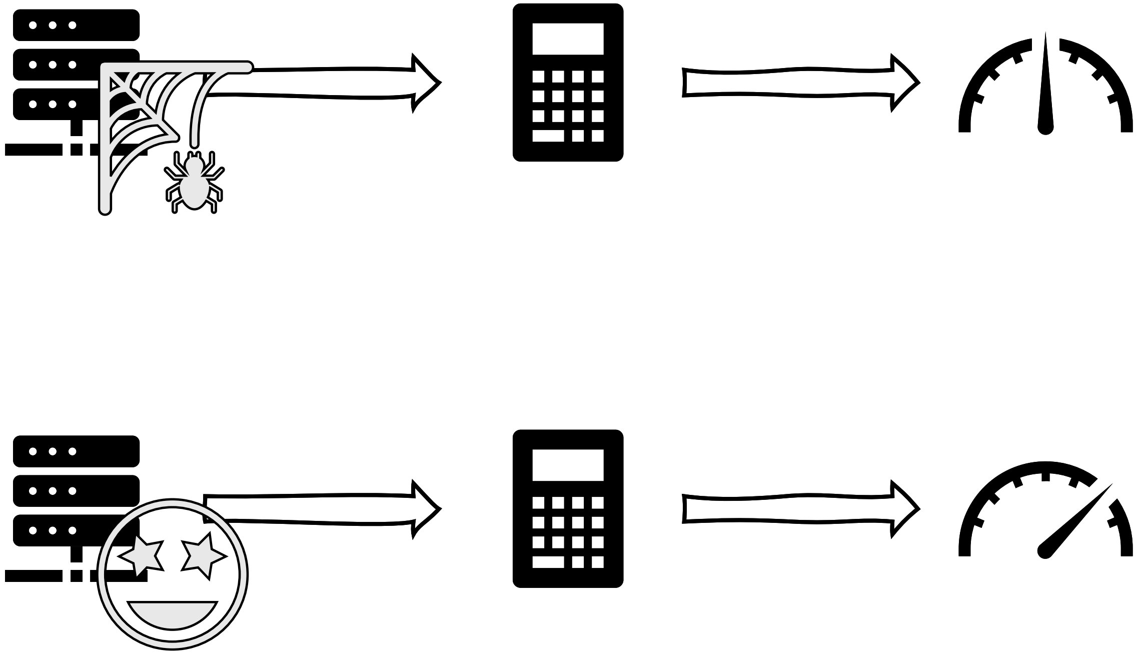
Rate each web page for how well it matches the query, using these numbers: 0 = no match, 1 = some match, 2 = great match. Think of writing a report on the query topic. A web page gets 2 if it is mainly about the topic or has important information for the report. A web page gets 1 if it has some information for the report, but also other stuff. A web page gets 0 if it has nothing to do with the topic or the report.

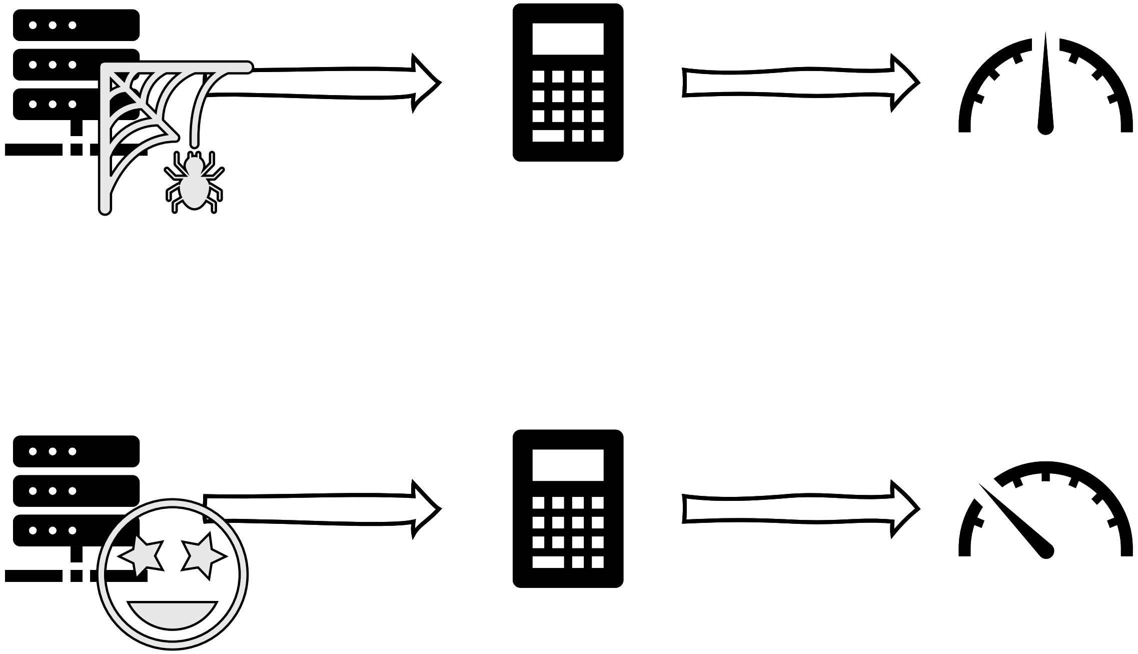
$\kappa = 0.72$

To rate a web page for a query, use 0, 1, or 2. Use 0 if the page has nothing to do with the query. Use 1 if the page has some useful information, but also other stuff. Use 2 if the page is mainly about the query or has important information.

$\kappa = 0.50$


1.00
0.75
0.50
0.25
0.00
κ against TREC assessors
All -DNA- paraphrases 95% interval
Original prompt -DNA- (best from Table 2)
Cormack et al.
Faggioli et al. best
Prompt ----- (basic)
Prompt R---M (worst from Table 2)





Regrettable decisions:
0/1000 prompt templates, 11/1000 paraphrases

At Bing

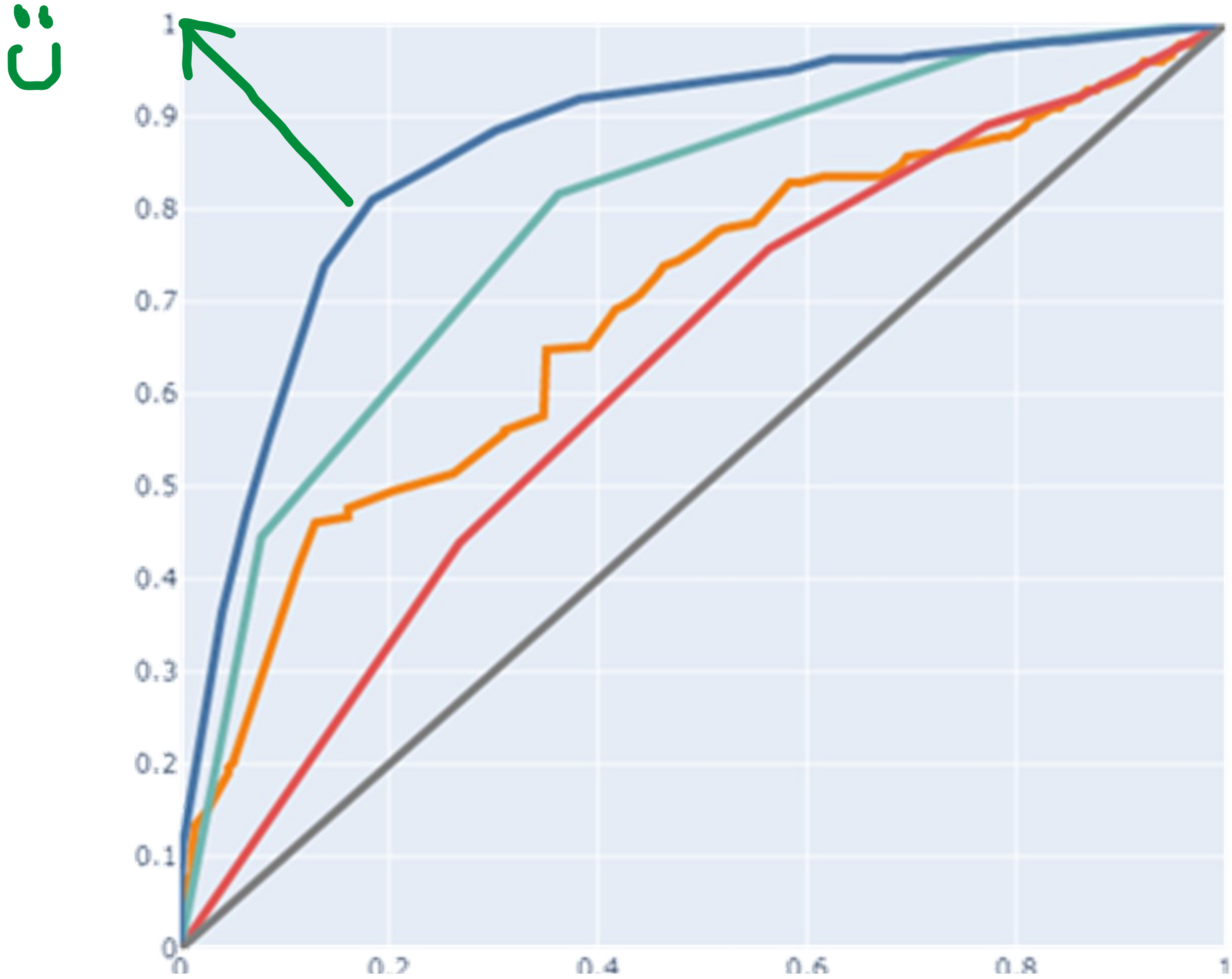
GPT-4

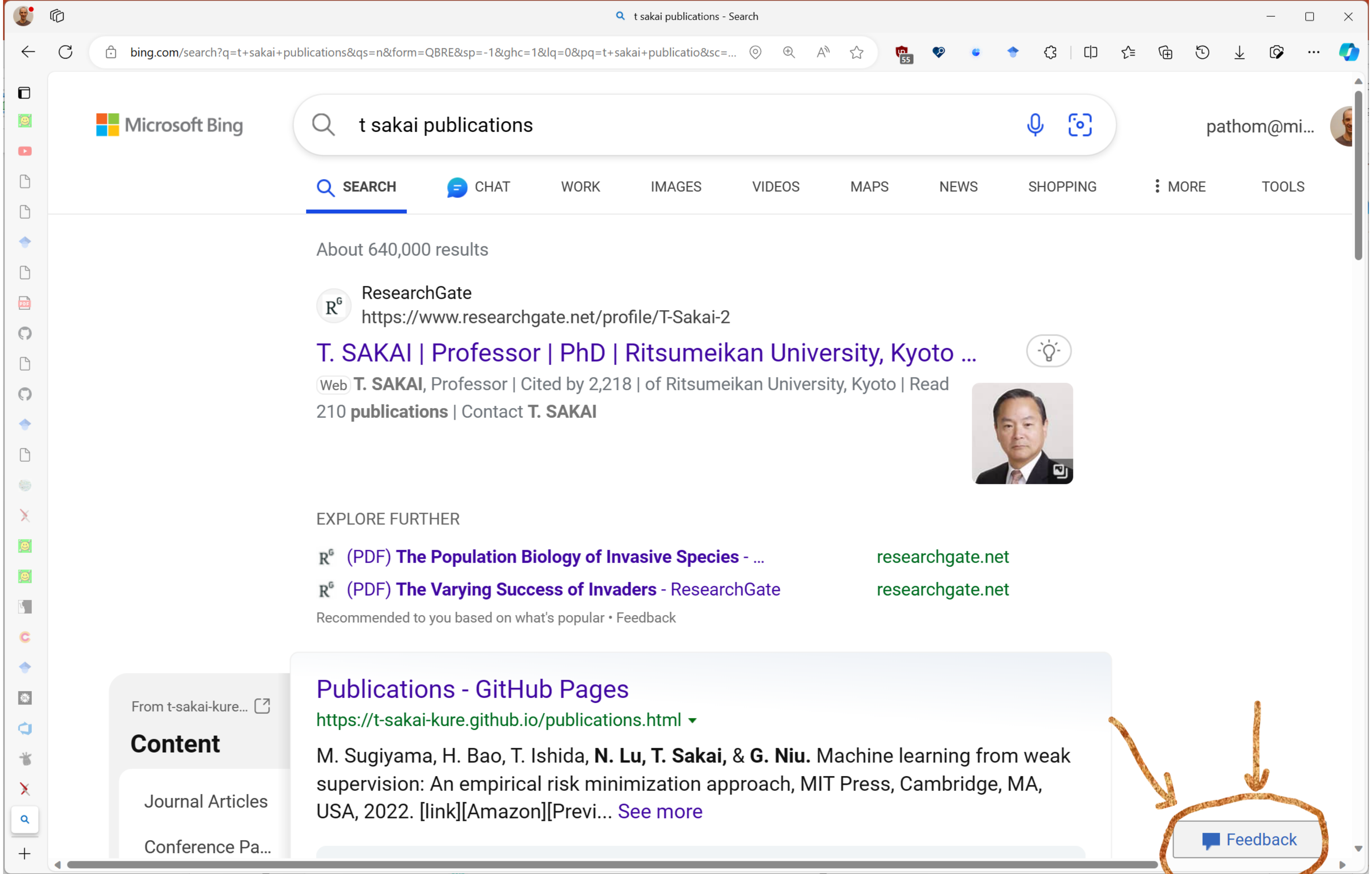
Best (trained) crowd

Trained crowd

Untrained crowd

Random





t sakai publications - Search
bing.com/search?q=t+sakai+publications&qs=n&form=QBRE&sp=-1&ghc=1&lq=0&pq=t+sakai+publicatio&sc=...
Microsoft Bing
pathom@mi...
SEARCH
CHAT
WORK
IMAGES
VIDEOS
MAPS
NEWS
SHOPPING
MORE
TOOLS

t sakai publications

About 640,000 results

ResearchGate
https://www.researchgate.net/profile/T-Sakai-2

T. SAKAI | Professor | PhD | Ritsumeikan University, Kyoto ...

Web **T. SAKAI**, Professor | Cited by 2,218 | of Ritsumeikan University, Kyoto | Read 210 **publications** | Contact **T. SAKAI**

EXPLORE FURTHER

(PDF) **The Population Biology of Invasive Species** - ... researchgate.net

(PDF) **The Varying Success of Invaders** - ResearchGate researchgate.net

Recommended to you based on what's popular • Feedback

From t-sakai-kure...

Content

Journal Articles

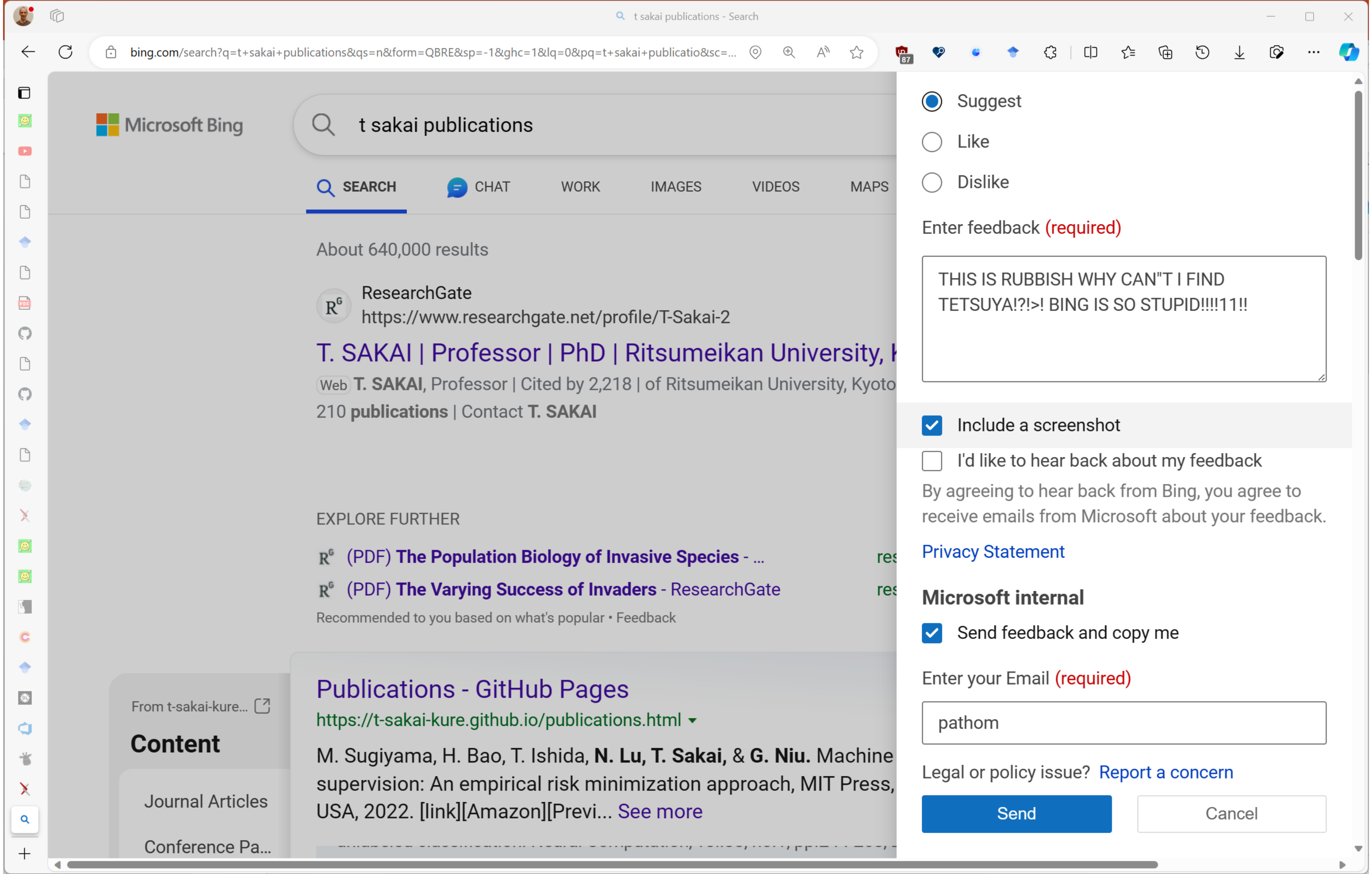
Conference Pa...

Publications - GitHub Pages

https://t-sakai-kure.github.io/publications.html

M. Sugiyama, H. Bao, T. Ishida, **N. Lu, T. Sakai,** & **G. Niu.** Machine learning from weak supervision: An empirical risk minimization approach, MIT Press, Cambridge, MA, USA, 2022. [link][Amazon][Previ... See more

Feedback



t sakai publications - Search

bing.com/search?q=t+sakai+publications&qs=n&form=QBRE&sp=-1&ghc=1&lq=0&pq=t+sakai+publicatio&sc=...

t sakai publications

About 640,000 results

ResearchGate
https://www.researchgate.net/profile/T-Sakai-2

T. SAKAI | Professor | PhD | Ritsumeikan University, K

Web **T. SAKAI**, Professor | Cited by 2,218 | of Ritsumeikan University, Kyoto
210 **publications** | Contact **T. SAKAI**

EXPLORE FURTHER

(PDF) **The Population Biology of Invasive Species** - ...

(PDF) **The Varying Success of Invaders** - ResearchGate

Recommended to you based on what's popular • Feedback

From t-sakai-kure...

Content

Journal Articles

Conference Pa...

Publications - GitHub Pages
https://t-sakai-kure.github.io/publications.html

M. Sugiyama, H. Bao, T. Ishida, **N. Lu, T. Sakai,** & **G. Niu.** Machine
supervision: An empirical risk minimization approach, MIT Press,
USA, 2022. [link][Amazon][Previ... See more

● Suggest

○ Like

○ Dislike

Enter feedback (required)

THIS IS RUBBISH WHY CAN"T I FIND TETSUYA!?!>! BING IS SO STUPID!!!!11!!

☑ Include a screenshot

☐ I'd like to hear back about my feedback

By agreeing to hear back from Bing, you agree to receive emails from Microsoft about your feedback.

Privacy Statement

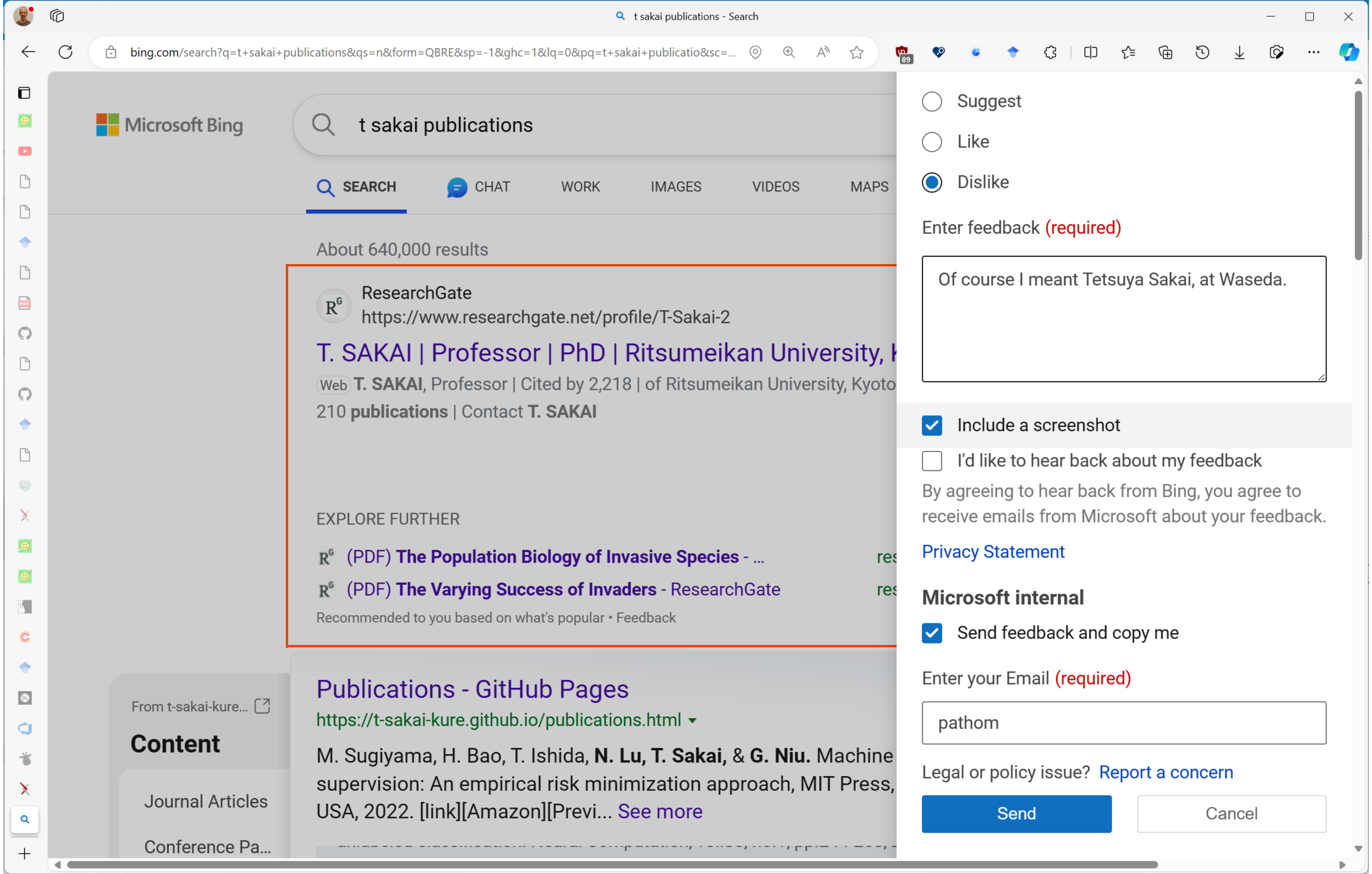
Microsoft internal

☑ Send feedback and copy me

Enter your Email (required)

pathom

Legal or policy issue? Report a concern



bing.com/search?q=t+sakai+publications&qs=n&form=QBRE&sp=-1&ghc=1&lq=0&pq=t+sakai+publicatio&sc=...

t sakai publications

About 640,000 results

ResearchGate
https://www.researchgate.net/profile/T-Sakai-2

T. SAKAI | Professor | PhD | Ritsumeikan University, K

Web **T. SAKAI**, Professor | Cited by 2,218 | of Ritsumeikan University, Kyoto
210 **publications** | Contact **T. SAKAI**

EXPLORE FURTHER

(PDF) **The Population Biology of Invasive Species** - ... res

(PDF) **The Varying Success of Invaders** - ResearchGate res

Recommended to you based on what's popular • Feedback

From t-sakai-kure...

Content

Journal Articles

Conference Pa...

Publications - GitHub Pages

https://t-sakai-kure.github.io/publications.html

M. Sugiyama, H. Bao, T. Ishida, **N. Lu, T. Sakai,** & **G. Niu.** Machine supervision: An empirical risk minimization approach, MIT Press, USA, 2022. [link][Amazon][Previ... See more

○ Suggest

○ Like

● Dislike

Enter feedback (required)

Of course I meant Tetsuya Sakai, at Waseda.

☑ Include a screenshot

□ I'd like to hear back about my feedback

By agreeing to hear back from Bing, you agree to receive emails from Microsoft about your feedback.

Privacy Statement

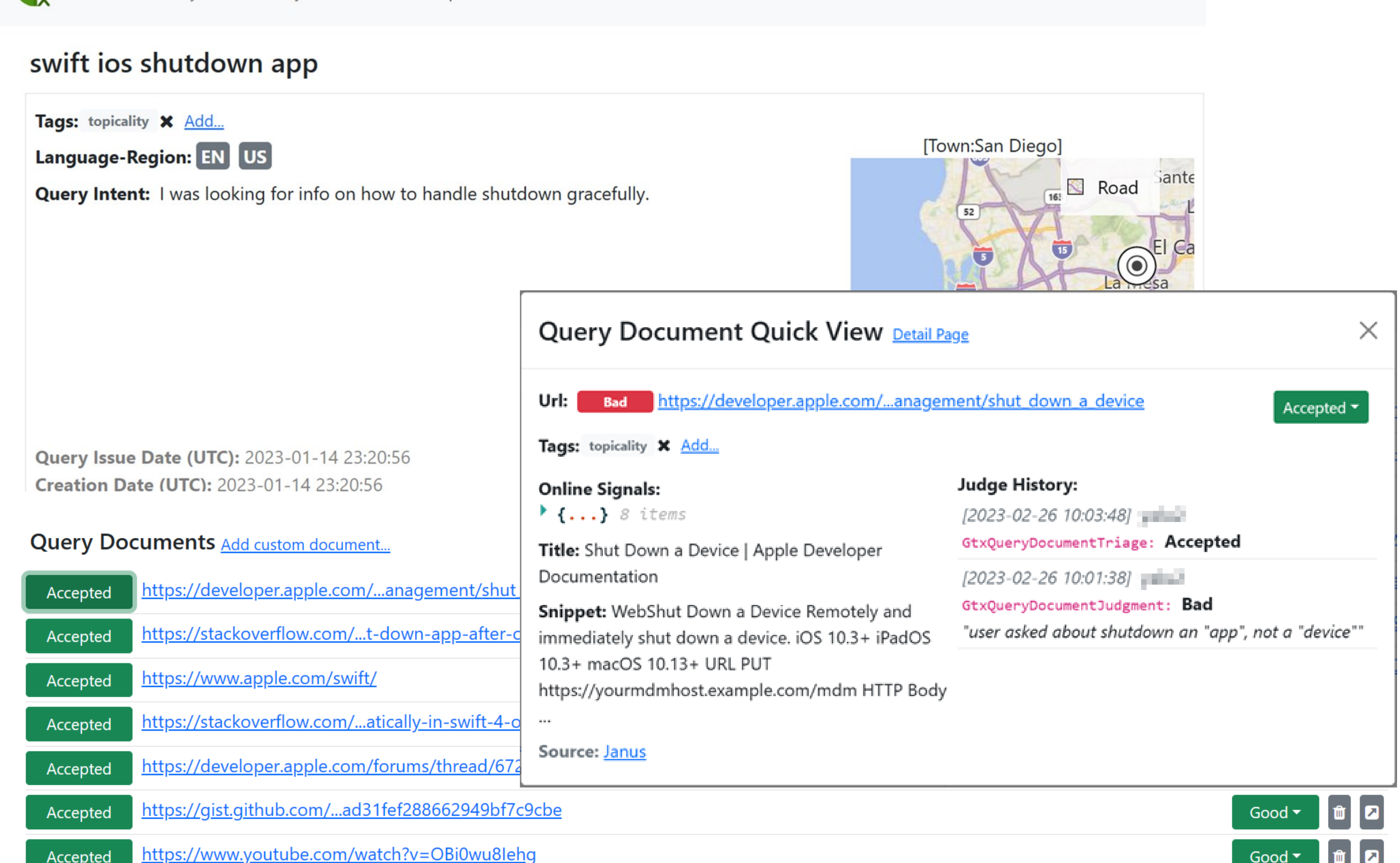
Microsoft internal

☑ Send feedback and copy me

Enter your Email (required)

pathom

Legal or policy issue? Report a concern



swift ios shutdown app

Tags: topicality ✖ Add...

Language-Region: EN US

Query Intent: I was looking for info on how to handle shutdown gracefully.

Query Issue Date (UTC): 2023-01-14 23:20:56

Creation Date (UTC): 2023-01-14 23:20:56

Query Documents

Add custom document...

Accepted https://developer.apple.com/...anagement/shut
Accepted https://stackoverflow.com/...t-down-app-after-c
Accepted https://www.apple.com/swift/
Accepted https://stackoverflow.com/...atically-in-swift-4-o
Accepted https://developer.apple.com/forums/thread/672
Accepted https://gist.github.com/...ad31fef288662949bf7c9cbe Good
Accepted https://www.youtube.com/watch?v=OBi0wu8lehg Good

Query Document Quick View

Detail Page

Url: Bad https://developer.apple.com/...anagement/shut_down_a_device Accepted

Tags: topicality ✖ Add...

Online Signals:

▸ {...} 8 items

Title: Shut Down a Device | Apple Developer Documentation

Snippet: WebShut Down a Device Remotely and immediately shut down a device. iOS 10.3+ iPadOS 10.3+ macOS 10.13+ URL PUT https://yourmdmhost.example.com/mdm HTTP Body ...

Source: Janus

Judge History:

[2023-02-26 10:03:48] [redacted]
`GtxQueryDocumentTriage:` **Accepted**

[2023-02-26 10:01:38] [redacted]
`GtxQueryDocumentJudgment:` **Bad**
"user asked about shutdown an "app", not a "device""

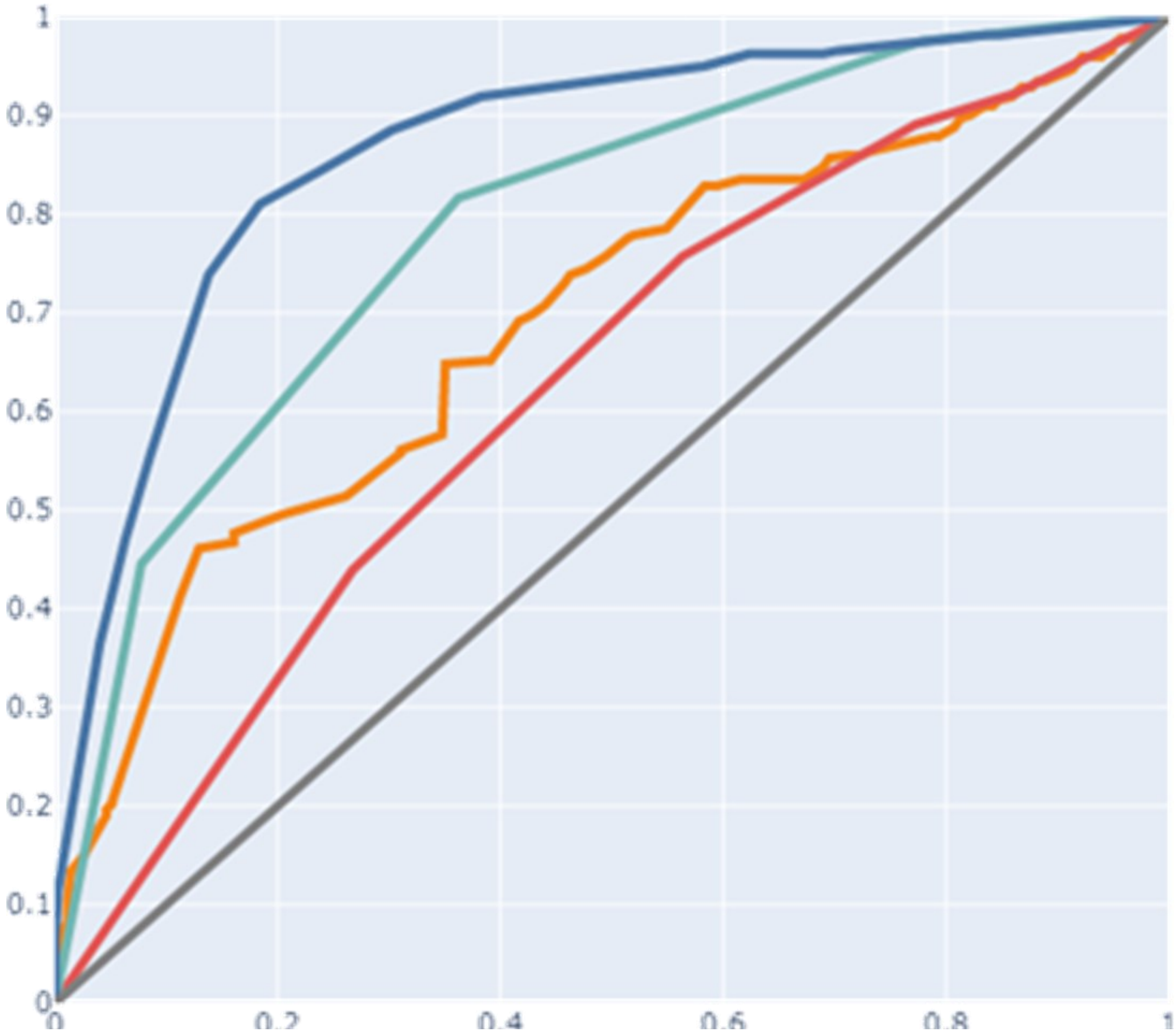
GPT-4

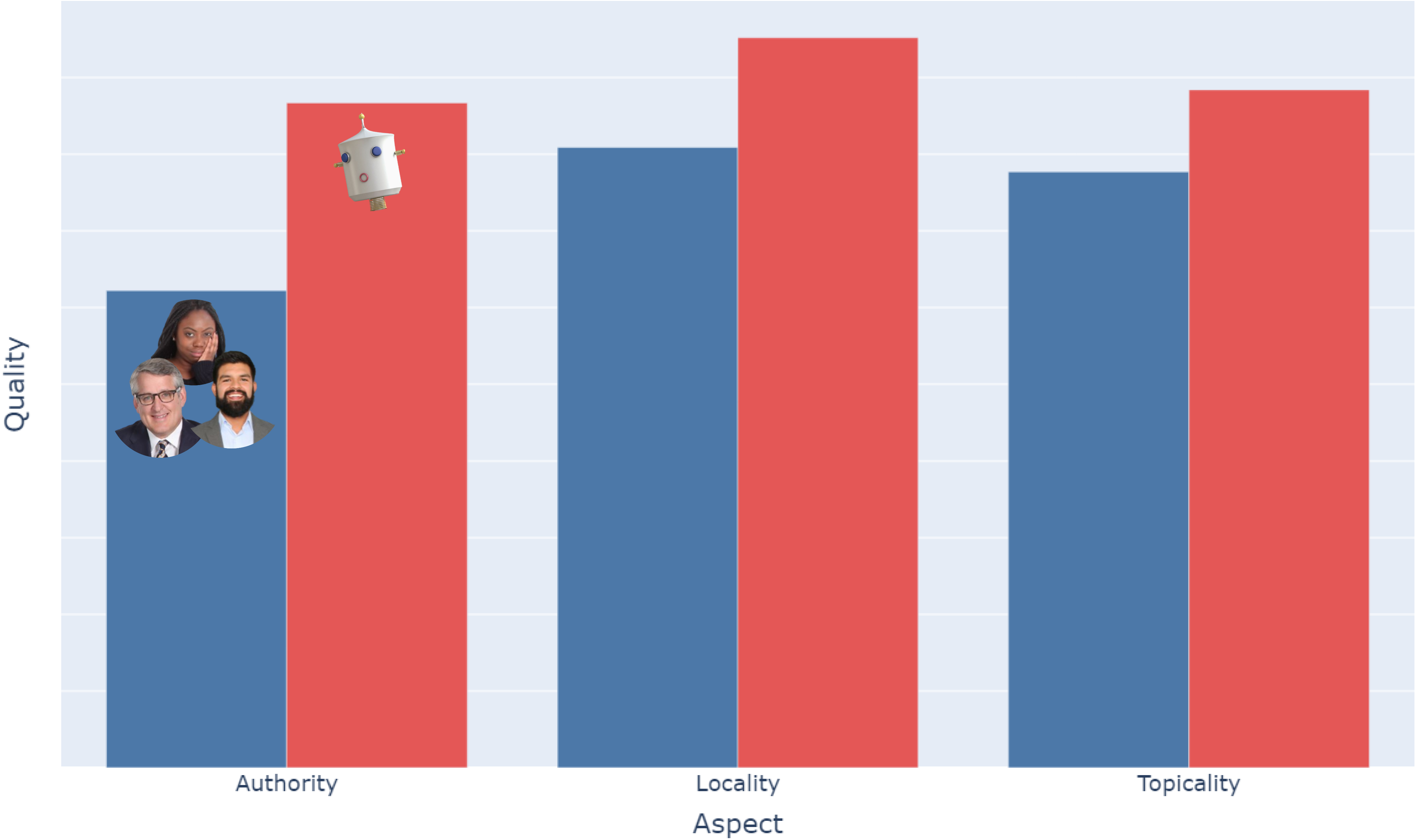
Best (trained) crowd

Trained crowd

Untrained crowd

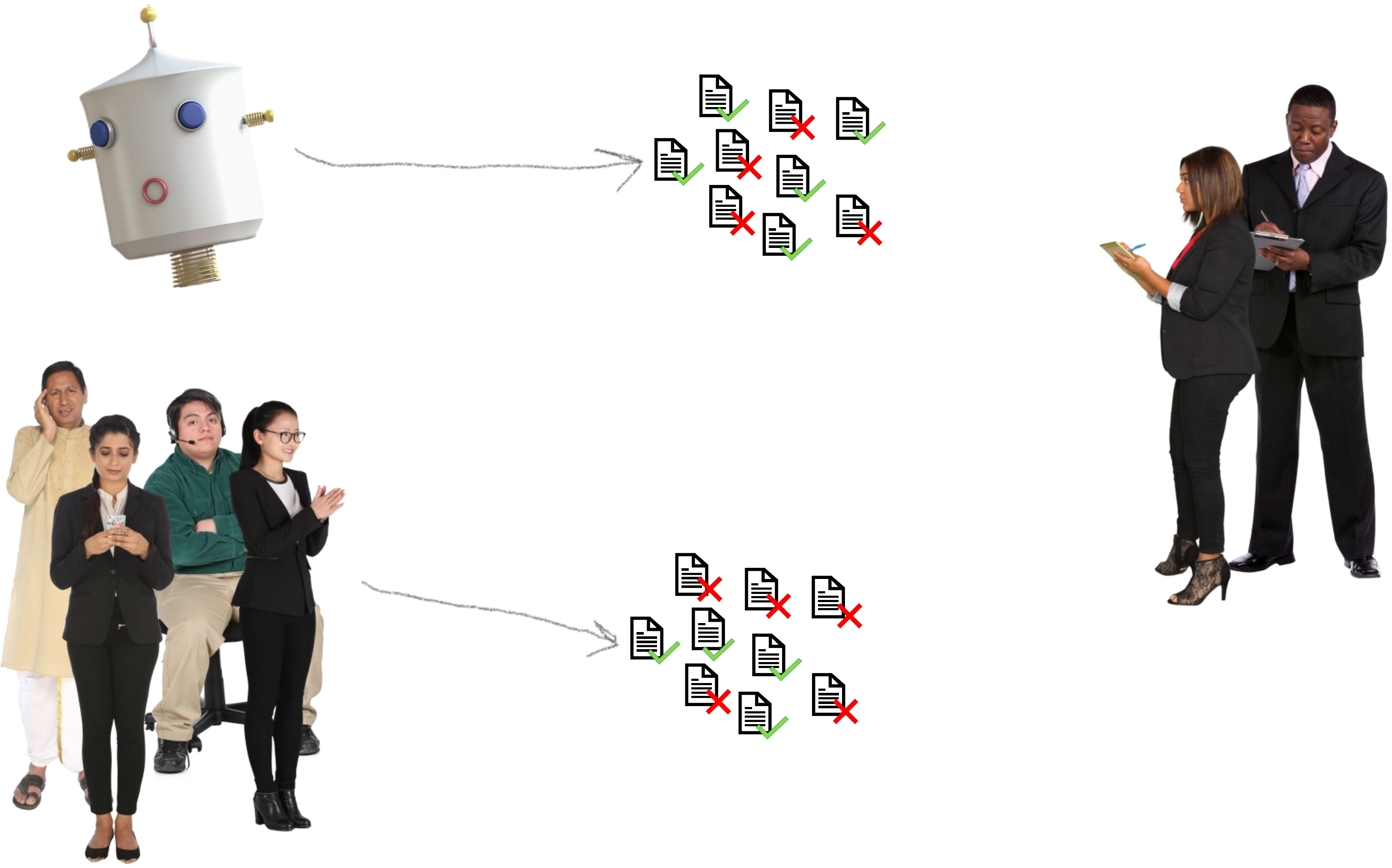
Random

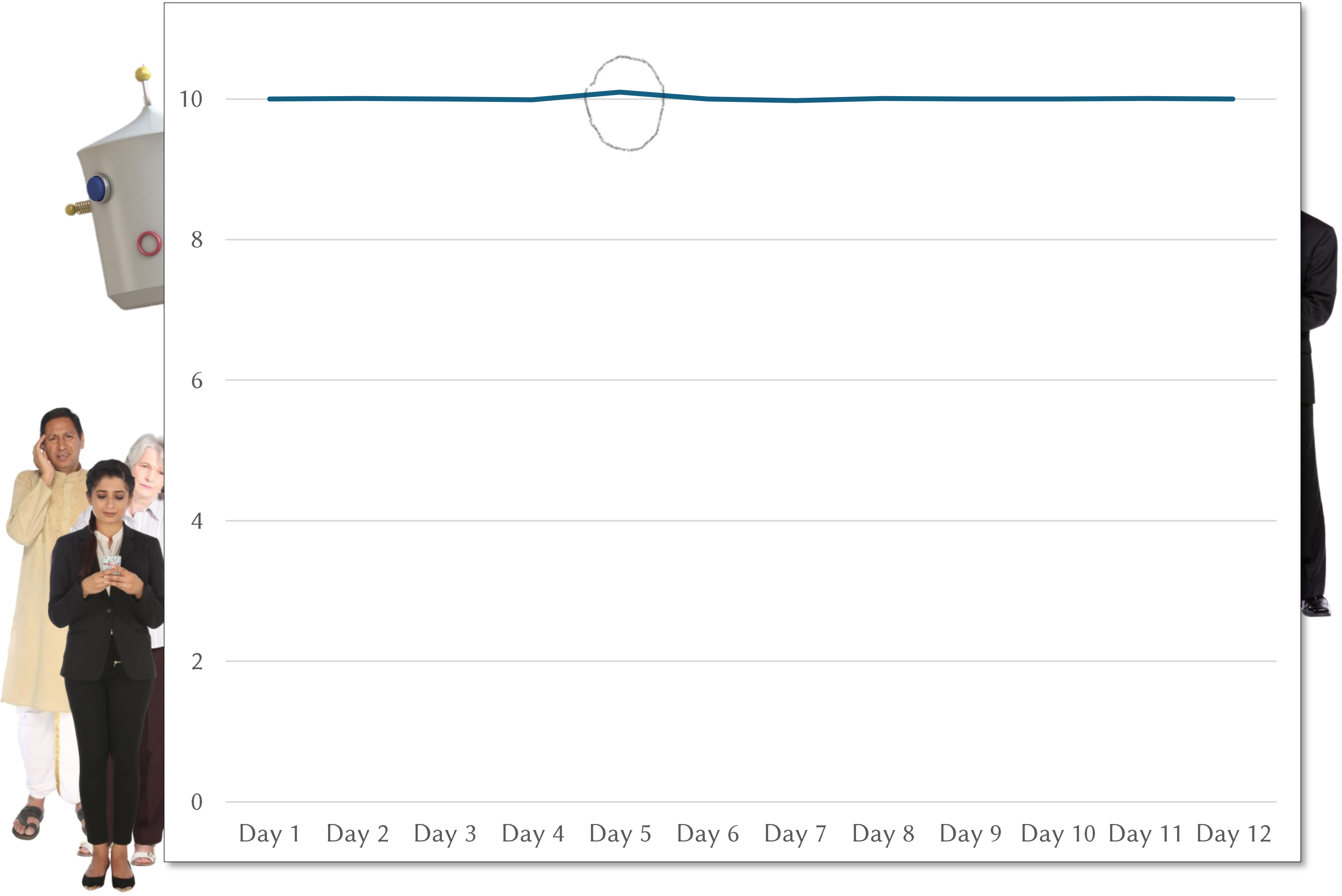



Quality
Authority
Locality
Topicality
Aspect

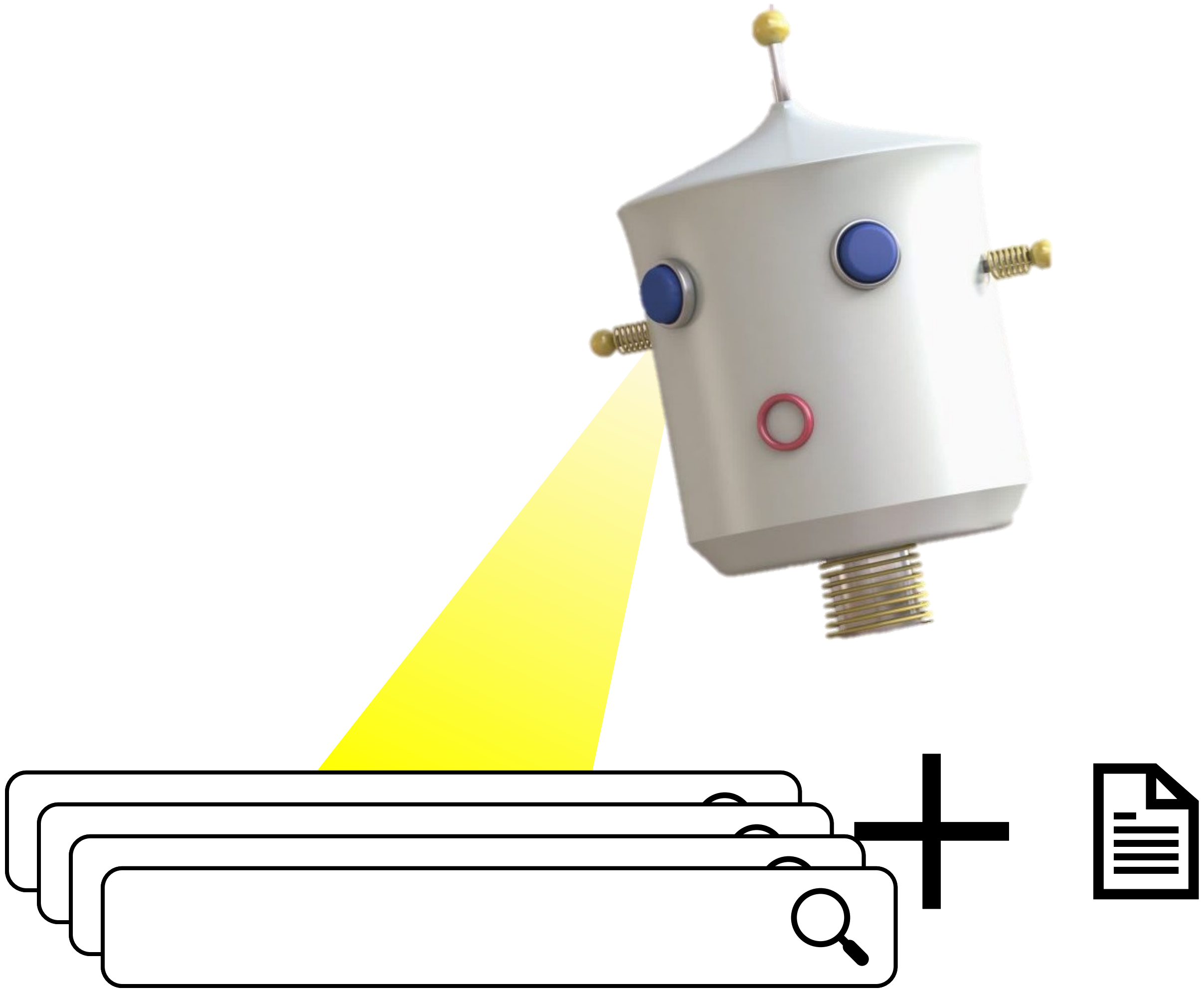
🎯 Motivation	Social & legal
📈 Correlation	Validity & fidelity
👍🏽 Agreement with gold standard	
💰 Cost	Efficiency
🚀 Throughput	
⏱️ Latency	
➡️ Direction & sensitivity to known changes	Reliability & sensitivity
😑 Stability	
🏏 Playing nicely with others	Organisational effects

	Relative	**Latency**	**Relative**	**Relative**
	accuracy		**throughput**	**cost**
Employees	+24%	Hours/days	$\times{}^{1}/_{100}$	×8
Best crowd	+19%	Hours/days	$\times{}^{1}/_{15}$	×5
Avg crowd	—	Hours	—	—
GPT-4	+28%	Mins/hours	×10	$\times{}^{1}/_{20}$

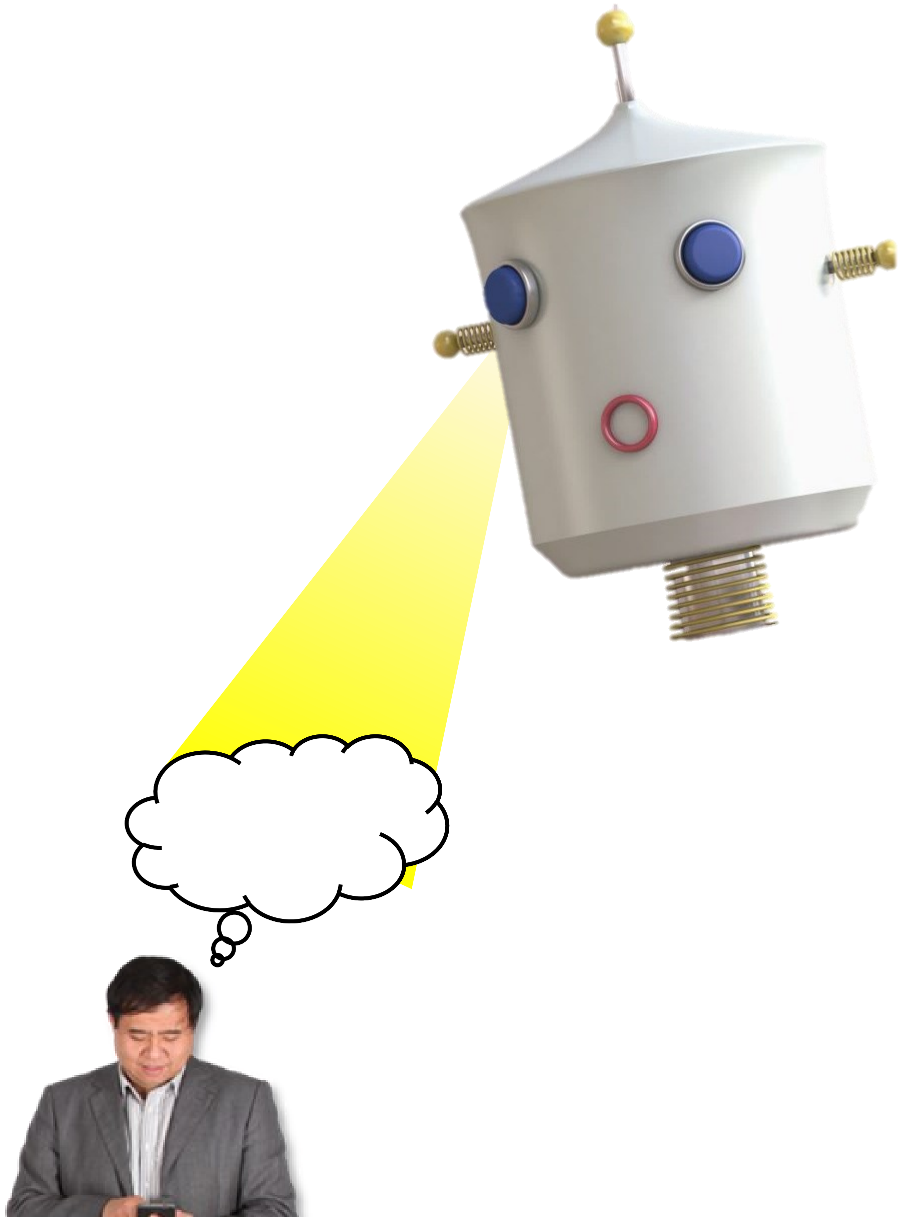


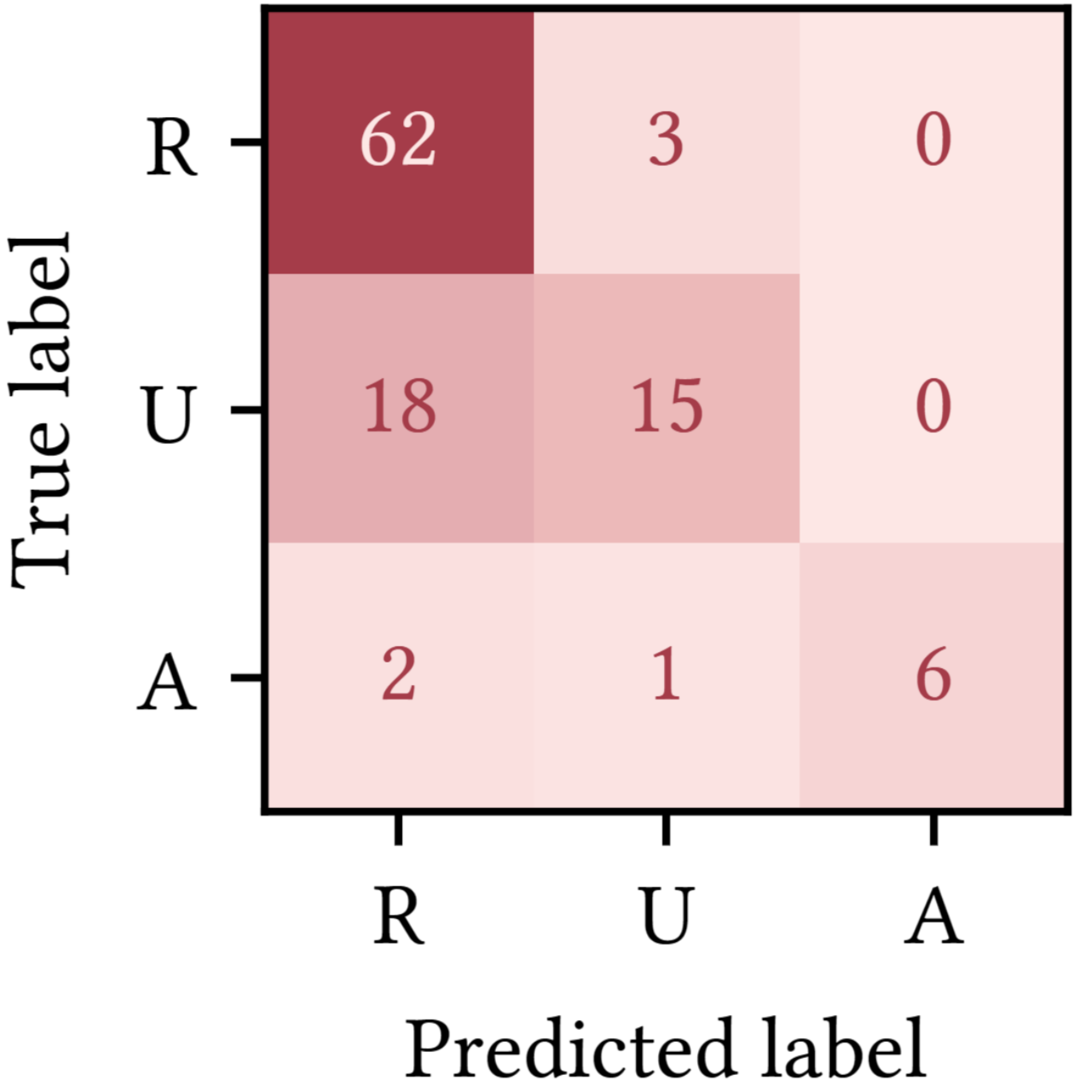

10
8
6
4
2
0
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
Day 8
Day 9
Day 10
Day 11
Day 12

Next steps



My cake es out why?
What is wrong with my cake?
Cake comes out flat reasons
Cake is not rising
Flat cake issue oven temp?
Causes of flat cake
Reasons for flat cake
Why won't my cake rise?





True label \ Predicted label	R	U	A
R	62	3	0
U	18	15	0
A	2	1	6

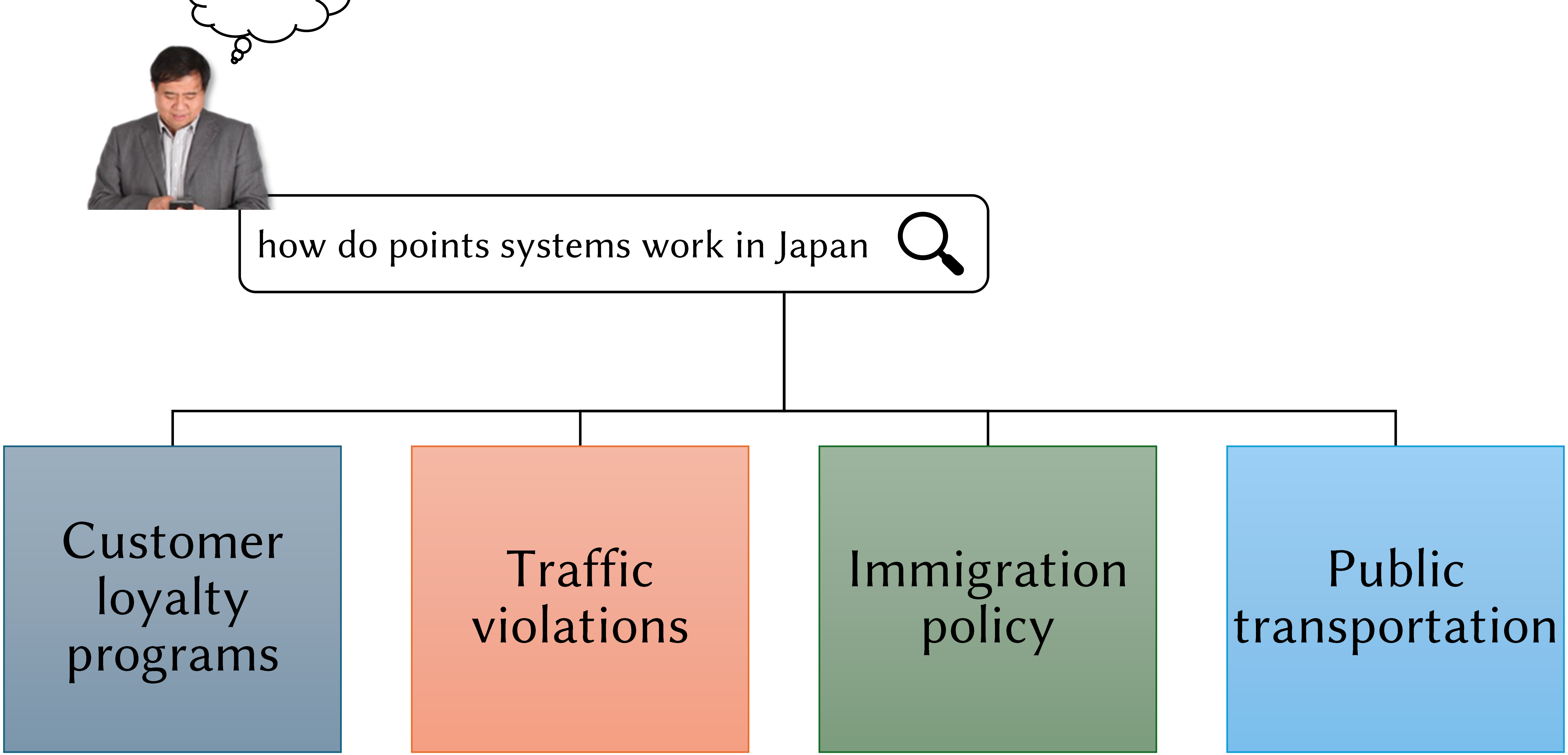
how do points systems work in Japan

Provide an explanation of how various loyalty card
programs work in Japan, including the be
requirements, and limitations of each. Inc
popular loyalty cards from different categories, such as
convenience stores, supermarkets, and re
comparison of the advantages and disad
loyalty cards versus other payment methods in Japan,
including current rewards and benefits. Highlight the most
popular services and participating merch

loyalty card programs Japan

best loyalty cards for travelers in Japan

managing loyalty points with phone apps



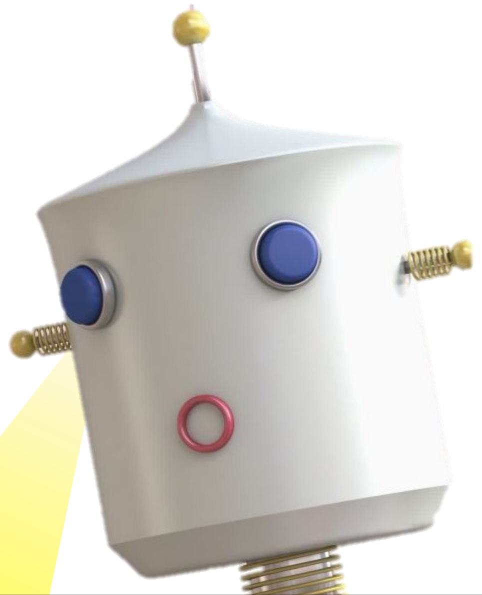
how do points systems work in Japan

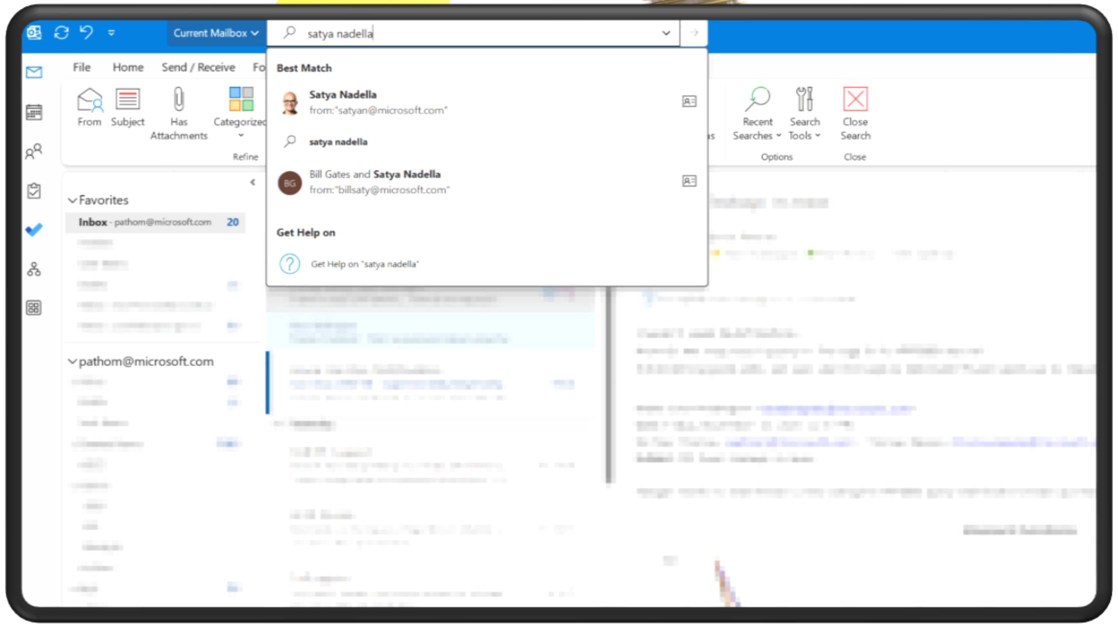
Customer loyalty programs

Traffic violations

Immigration policy

Public transportation





Best Match

Satya Nadella
from:"satyan@microsoft.com"

satya nadella

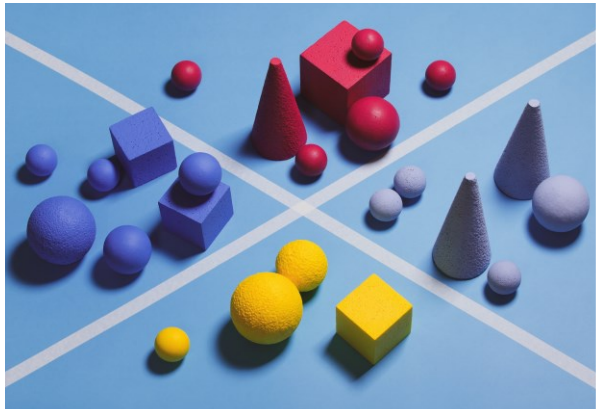
Bill Gates and Satya Nadella
from:"billsaty@microsoft.com"

Get Help on

Get Help on "satya nadella"

Onward

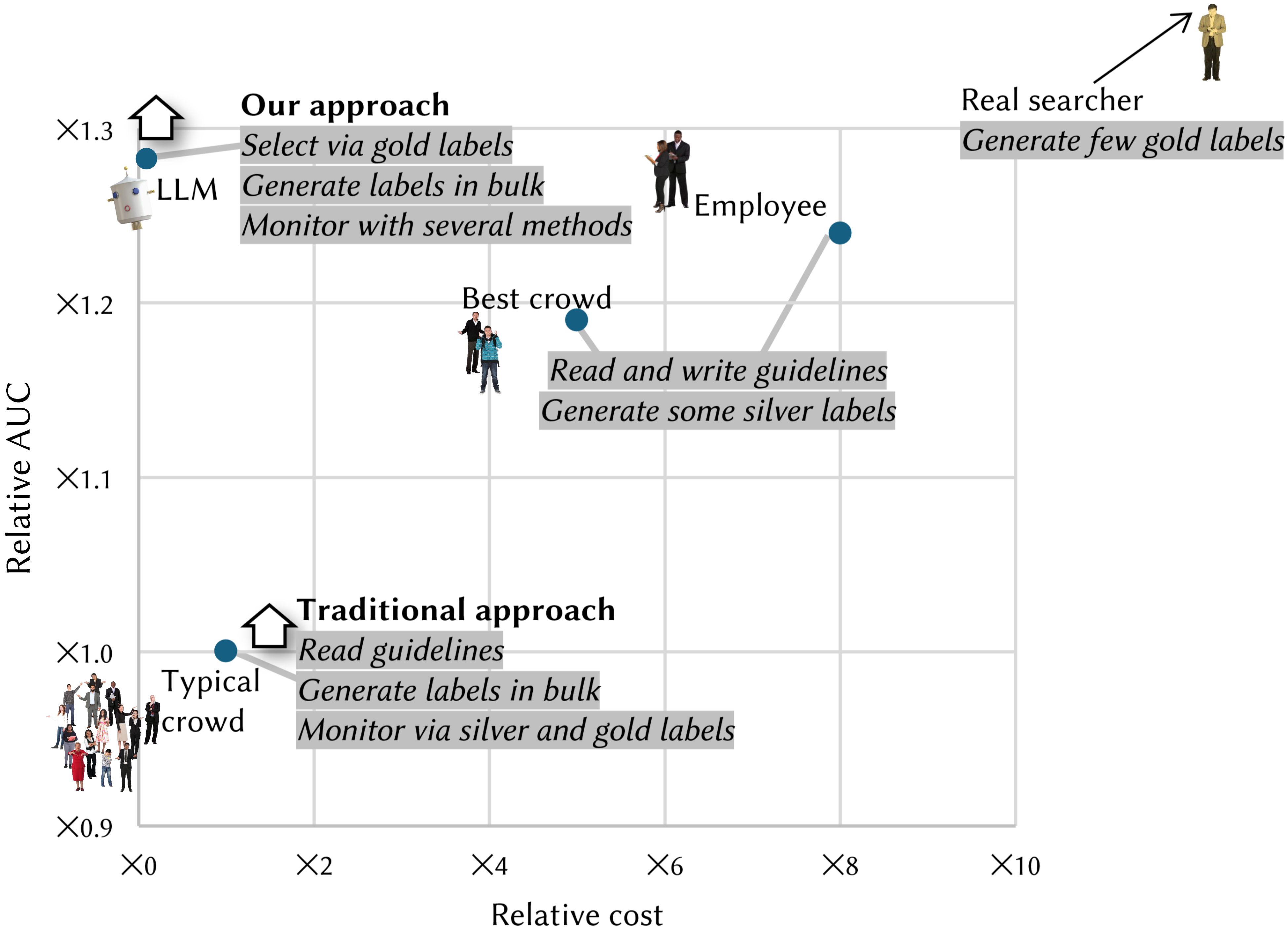
Bias



Over-fitting

Optimisation

Cost


Our approach
Select via gold labels
Generate labels in bulk
Monitor with several methods
LLM
Real searcher
Generate few gold labels
Employee
Best crowd
Read and write guidelines
Generate some silver labels
Traditional approach
Read guidelines
Generate labels in bulk
Monitor via silver and gold labels
Typical
crowd
Relative AUC
×1.3
×1.2
×1.1
×1.0
×0.9
×0
×2
×4
×6
×8
×10
Relative cost

👍 **We can use LLMs for labelling relevance;**
≈ TREC judges, > crowd.

🤔 **Many choices make a difference;**
so we need meta-metrics and audits.

🪙 **True “gold” judgements** make it possible to experiment.

📈 We’ve found LLMs very productive.

https://arxiv.org/abs/2309.10621