

Personalization and Conversation (and GPT-4)

Nick Craswell
Microsoft, USA



Disclaimer

This talk does not cover technologies used in Bing Chat or BizChat.

Instead, the content is: 1) Published research, and 2) Toy GPT-4 (dv3) examples to help us think through the implications of LLMs. I would enjoy feedback and discussion!

Topic

Published Work

GPT-4 Toy Examples

1. Deep learning in search (MS MARCO)

- Bajaj, Campos, Craswell, Deng, Gao, Liu, Majumder, McNamara, Mitra, Nguyen, Rosenberg. *MS MARCO: A human generated machine reading comprehension dataset*. arXiv preprint arXiv:1611.09268, 2018.
- Do we need more large datasets like MS MARCO? Synthetic data.

2. Personalized search

- Does GPT-4 “understand” the dimensions of personalization? Synthetic data.

3. Conversational search

- Radlinski and Craswell. *A theoretical framework for conversational search*. CHIIR 2017
- Gao, Xiong, Bennett, Craswell. *Neural approaches to conversational information retrieval*. Vol. 44. Springer Nature, 2023
- Does GPT-4 conform to our expectations in conversational use?

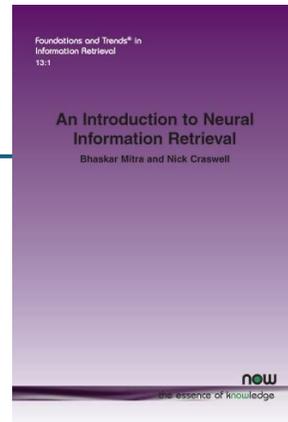
Deep Learning in Search (MS MARCO)

MS MARCO helped, but do we need another one?

1990s: BM25. 2000s: Learning to rank.

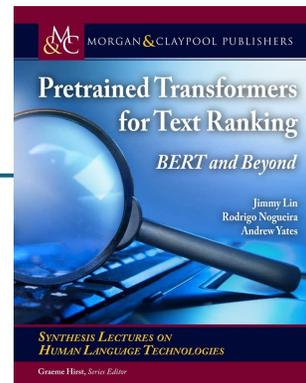
Public data:
~100 queries.

Proprietary data:
~100k queries



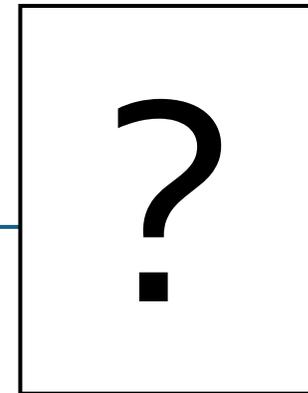
~2019: MS MARCO, BERT, and more.

MS MARCO: Public
100+k queries.
BERT much better
than BM25.



~2022: ChatGPT. BEIR.

BEIR: Beyond MS
MARCO data.
Synthetic data.

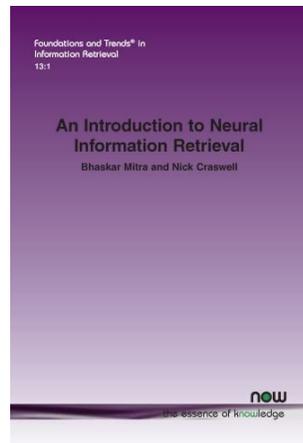


~2019 turning point

2018-10-25: BERT

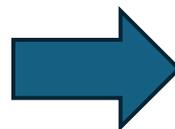
2018-10-26: MS MARCO Ranking

2019-01-07: BERT tops MS MARCO Ranking leaderboard



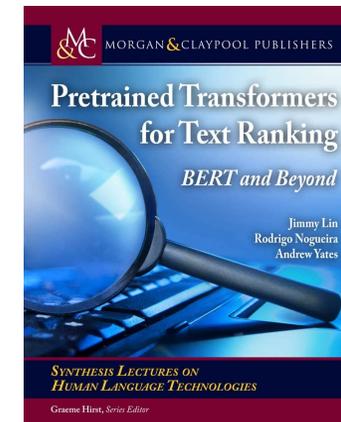
Before

Dataset: robusto4
0.5M docs
0 train queries
250 test queries



After

Dataset: MS MARCO
8.8M passages
400k train queries
7k dev queries
~50 new test q per year



Armstrong, Moffat, Webber, and Zobel. "Improvements that don't add up: ad-hoc retrieval results since 1998." 2009

Mitra and Craswell. "An introduction to neural information retrieval." FNTIR 2018

Lin. "The neural hype and comparisons against weak baselines." SIGIR Forum 2019

Craswell, Mitra, Yilmaz, Campos, and Voorhees. "Overview of the TREC 2019 deep learning track." TREC 2019

Nogueira, and Cho. "Passage Re-ranking with BERT." *arXiv preprint arXiv:1901.04085* (2019).

Lin, Nogueira, and Yates. "Pretrained transformers for text ranking: Bert and beyond." Book 2021

Lin. "The neural hype, justified! A recantation." SIGIR Forum, 2021

MS MARCO: A Human Generated MACHine Reading COmprehension Dataset

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary,
Rangan Majumder and Li Deng
Microsoft AI & Research
Bellevue, WA, USA

{trnguye, miriamr, xiaso, jfgao, satiwary, ranganm, deng}@microsoft.com

2016: MS MARCO NLG dataset (100k Q).

MS MARCO: A Human Generated MACHine Reading COmprehension Dataset

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao,
Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen,
Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang
Microsoft AI & Research

2018: Expanded NLG task to 1M Q.
New idea: Propose a ranking task too!
Ranking leaderboard in late 2018...

Submission history

From: Bhaskar Mitra [\[view email\]](#)

[v1] Mon, 28 Nov 2016 18:14:11 UTC (1,852 KB)

[v2] Tue, 29 Nov 2016 02:39:53 UTC (1,852 KB)

[v3] Wed, 31 Oct 2018 14:46:47 UTC (579 KB)

[\[PDF\] MS MARCO: A human generated machine reading comprehension dataset](#)

[T.Nguyen](#), [M Rosenberg](#), [X Song](#), [J Gao](#), [S Tiwary](#)... - choice, 2016 - ceur-ws.org

... MACHine Reading Comprehension (**MS MARCO**) - a large ... a sample of those in **MS MARCO**.
We believe a characteristic of ... **MS MARCO** includes 100,000 questions, 1 million passages, ...

★ Save [🔗](#) Cite Cited by 1100 Related articles All 4 versions [🔗](#)

[Ms marco: A human generated machine reading comprehension dataset](#)

[P Bajaj](#), [D Campos](#), [N Craswell](#), [L Deng](#), [J Gao](#)... - arXiv preprint arXiv ..., 2016 - arxiv.org

... **MS MARCO** presents a challenging dataset for benchmarking MRC models. The **MS MARCO**
... Finally, we summarize the key distinguishing features of the **MS MARCO** dataset as follows...

☆ Save [🔗](#) Cite Cited by 423 Related articles All 2 versions [🔗](#)

BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

[NeurIPS 2021]

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych
 Ubiquitous Knowledge Processing Lab (UKP Lab)
 Department of Computer Science, Technical University of Darmstadt

If you have 100k training queries in-domain, great!
 BEIR: What if you don't have in-domain training Q?

Model	Model Name	In-domain test	
		MSMARCO	Average
BM25-Anserini	Anserini (Default Lucene, k=0.9, b=0.4)	0	0
BM25	Elasticsearch (Default Settings)	-0.01	-0.011
DistilBERT (TAS-B)	msmarco-distilbert-base-tas-b	0.18	-0.02
BM25-Anserini-RM3	Anserini (k=0.9, b=0.4) (RM3: 10,10,0.5)	0.16	-0.044
ANCE	msmarco-roberta-base-ance-fristp	0.16	-0.047
DistilBERT	msmarco-distilbert-base-v3	0.161	-0.054
DistilBERT (dot)	msmarco-distilbert-base-dot-prod-v3	0.161	-0.061
MiniLM-L-12	msmarco-MiniLM-L-12-v3	0.157	-0.08
MiniLM-L-6	msmarco-MiniLM-L-6-v3	0.151	-0.088
DistilBERT (NQ)	nq-distilbert-base-v1	0.013	-0.164
DPR (Multi)	facebook-dpr-ctx_encoder-multiset-base	-0.051	-0.174
DPR (KILT)	to be released by facebook	-0.078	-0.212

Model	Model Name	In-domain test		non-MS MARCO test sets																
		MSMARCO	Average	TREC-COV	BIOASQ	NFCorpus	NQ	HotpotQA	FIQA	Signal-1M	TREC-NEW	ArguAna	Touche-2C	DBpedia	SCIDOCS	FEVER	Climate-FI	SciFact	Robust04	
BM25-Anserini	Anserini (Default Lucene, k=0.9, b=0.4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BM25	Elasticsearch (Default Settings)	-0.01	-0.011	-0.04	0.049	-0.028	-0.019	-0.002	0.003	0.058	-0.027	0.027	-0.367	-0.025	-0.002	-0.105	-0.034	-0.045	-0.021	-0.021
DistilBERT (TAS-B)	msmarco-distilbert-base-tas-b	0.18	-0.02	-0.175	-0.082	-0.006	0.134	-0.019	0.064	0.041	-0.021	0.013	-0.194	0.071	-0.009	-0.053	0.015	-0.022	0.019	0.019
BM25-Anserini-RM3	Anserini (k=0.9, b=0.4) (RM3: 10,10,0.5)	0.16	-0.044	-0.002	-0.035	0.023	-0.074	-0.147	-0.043	-0.009	0.035	-0.164	-0.367	-0.04	-0.011	-0.206	-0.029	-0.023	0.018	0.018
ANCE	msmarco-roberta-base-ance-fristp	0.16	-0.047	-0.002	-0.159	-0.088	0.117	-0.147	0.059	-0.081	-0.016	0.001	-0.083	-0.032	-0.036	-0.084	-0.015	-0.158	-0.016	-0.016
DistilBERT	msmarco-distilbert-base-v3	0.161	-0.054	-0.074	-0.17	-0.068	0.121	-0.109	0.022	-0.069	-0.031	0.015	-0.118	0.026	-0.025	-0.083	-0.008	-0.134	-0.071	-0.071
DistilBERT (dot)	msmarco-distilbert-base-dot-prod-v3	0.161	-0.061	-0.021	-0.185	-0.058	0.143	-0.126	0.019	-0.087	-0.041	-0.073	-0.07	0.002	-0.045	-0.068	-0.044	-0.154	-0.097	-0.097
MiniLM-L-12	msmarco-MiniLM-L-12-v3	0.157	-0.08	-0.183	-0.186	-0.074	0.093	-0.147	0.004	-0.059	-0.038	-0.007	-0.142	-0.006	-0.045	-0.182	-0.034	-0.162	-0.113	-0.113
MiniLM-L-6	msmarco-MiniLM-L-6-v3	0.151	-0.088	-0.177	-0.209	-0.107	0.065	-0.155	-0.005	-0.071	-0.056	-0.02	-0.144	-0.021	-0.042	-0.158	-0.048	-0.17	-0.115	-0.115
DistilBERT (NQ)	nq-distilbert-base-v1	0.013	-0.164	-0.263	-0.312	-0.149	0.255	-0.298	-0.083	-0.169	-0.159	-0.104	-0.109	-0.106	-0.093	-0.454	-0.089	-0.277	-0.202	-0.202
DPR (Multi)	facebook-dpr-ctx_encoder-multiset-base	-0.051	-0.174	-0.324	-0.338	-0.136	0.145	-0.212	-0.124	-0.175	-0.237	-0.239	-0.24	-0.05	-0.081	-0.191	-0.065	-0.347	-0.156	-0.156
DPR (KILT)	to be released by facebook	-0.078	-0.212	-0.479	-0.397	-0.194	0.131	-0.085	-0.197	-0.203	-0.288	-0.389	-0.367	-0.027	-0.108	-0.067	-0.017	-0.546	-0.357	-0.357

Latest BEIR leaderboard

Process corpus, e.g. generate training data.

Participant team	Avg
nle ((BM25+SPLADE) RANKT5 top 50)	0.124
ZA+NM+Unicamp (InParsv2)	0.117
MetaAI+UW*2 (DRAGON+)	0.045
BEIR (SPLADE)	0.045
BEIR (BM25 multifield)	0
BEIR (UCx)	-0.001
BEIR (BM25 flat)	-0.005
BEIR (BM25 flat-wp)	-0.025

Participant team	Avg	TREC-COV	BioASQ	NFCorpus	NQ	HotpotQA	FiQA	Signal-1M	TREC-NEWS	Robust04	ArguAna	Touche-2C	COADupSI	Quora	DBPedia	SCIDOCS	FEVER	Climate-FI	SciFact	Last submission	
nle ((BM25+SPLADE) RANKT5 top 50)	0.124	0.183	0.128	0.057	0.308	0.166	0.215	0.007	0.122	0.189	0.14	-0.006	0.113	0.103	0.177	0.03	0.116	0.091	0.087	2 months ago	
ZA+NM+Unicamp (InParsv2)	0.117	0.19	0.13	0.06	0.309	0.188	0.273	-0.022	0.092	0.225	-0.045	-0.076	0.149	0.056	0.185	0.05	0.119	0.11	0.109	3 months ago	
MetaAI+UW*2 (DRAGON+)	0.045	0.103	-0.032	0.014	0.208	0.059	0.12	-0.029	0.046	0.072	0.055	-0.104	0.055	0.086	0.104	0.001	0.028	0.014	0.014	2 months ago	
BEIR (SPLADE)	0.045	0.055	0.039	0.02	0.215	0.083	0.115	-0.034	-0.004	0.051	0.107	-0.123	0.042	0.025	0.129	0.001	0.043	0.015	0.034	4 months ago	
BEIR (BM25 multifield)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4 months ago
BEIR (UCx)	-0.001	-0.016	0.012	-0.008	0.096	0.064	0.053	-0.055	-0.024	-0.004	-0.018	-0.069	0.002	-0.126	0.025	-0.014	0.059	-0.031	0.021	4 months ago	
BEIR (BM25 flat)	-0.005	-0.061	-0.058	-0.003	-0.023	0.03	0	0	-0.003	0	-0.017	0.075	0.003	0	0.005	-0.009	-0.102	-0.048	0.014	4 months ago	
BEIR (BM25 flat-wp)	-0.025	-0.091	-0.046	-0.011	-0.024	-0.01	-0.018	0.02	-0.037	-0.03	-0.05	0.099	-0.004	-0.059	-0.029	-0.02	-0.095	-0.055	0.007	4 months ago	

Holistic Evaluation of Language Models

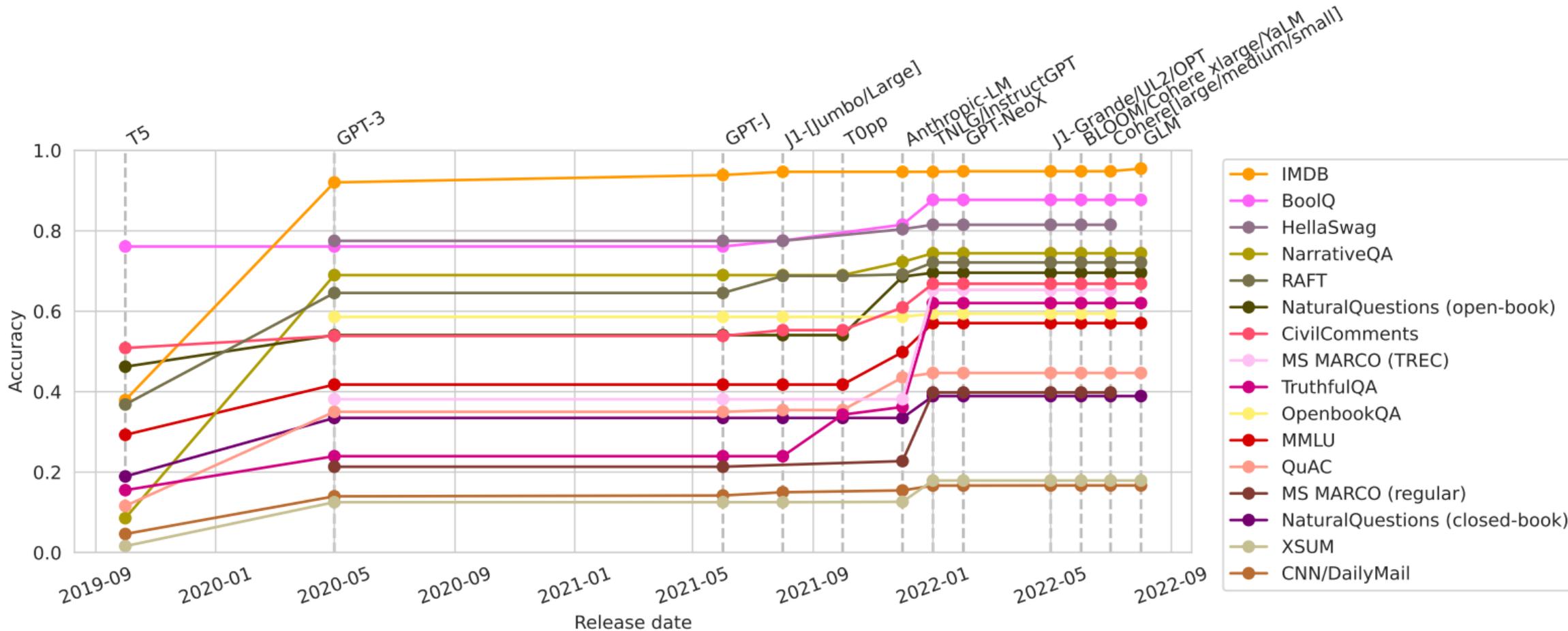
Percy Liang[†] Rishi Bommasani[†] Tony Lee^{†1}

Dimitris Tsipras* Dilara Soylu* Michihiro Yasunaga* Yian Zhang* Deepak Narayanan* Yuhuai Wu*²

Ananya Kumar Benjamin Newman Binhang Yuan Bobby Yan Ce Zhang

Drew
Hua
Mira

Tat

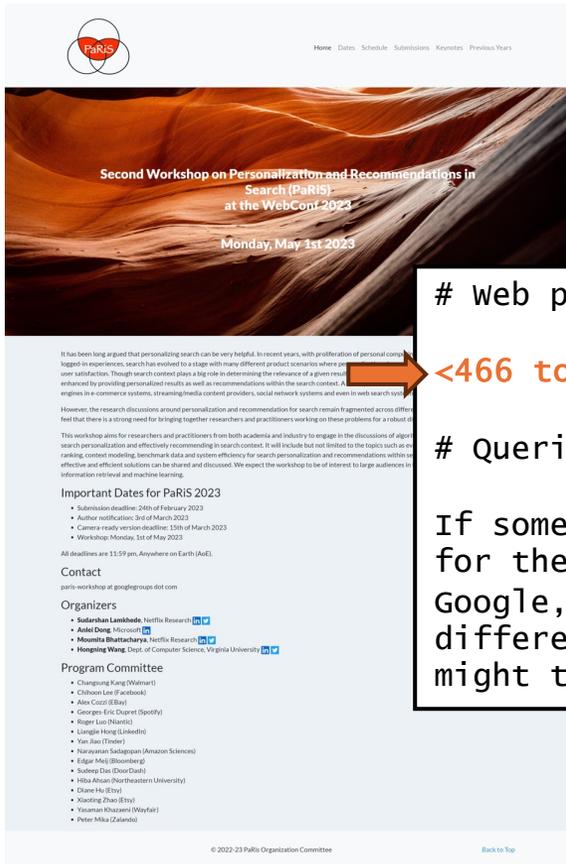


Cohere / Cohere medium v20220720 (6.1B)
 Cohere / Cohere small v20220720 (410M)
 Cohere / Cohere xlarge v20221108 (52.4B)
 Cohere / Cohere medium v20221108 (6.1B)

TruthfulQA
 Information retrieval
 MS MARCO (regular)
 MS MARCO (TREC)

• HellaSwag
 • OpenbookQA
 • TruthfulQA
 • MMLU
 • BoolQ
 • LegalSupport
 • LSIAT
 • BBQ
 Vary prompts
 • NaturalQuestions (open-book)
 • CNN/DailyMail
 • IMDB
 • CivilComments
 Robustness to context sets
 • IMDB
 • BoolQ

• BLEUScore (F1)
 • Coverage
 • Density
 • Compression
 • HumanEval-factfulness
 • HumanEval-relevance
 • HumanEval-coherence
 APPS metrics
 • Avg # tests passed
 • Strict correctness
 BBQ metrics
 • BBQ (ambigous)
 • BBQ (unambiguous)
 Copyright metrics
 • Longest common prefix length
 • Edit distance (Levenshtein)
 • Edit similarity (Levenshtein)
 Discrimination metrics
 • Self-BLEU
 • Entropy (Monte Carlo)



Toy GPT-4 Example: Generate a (bad) MS MARCO dataset.

Web page content

<466 tokens were here>

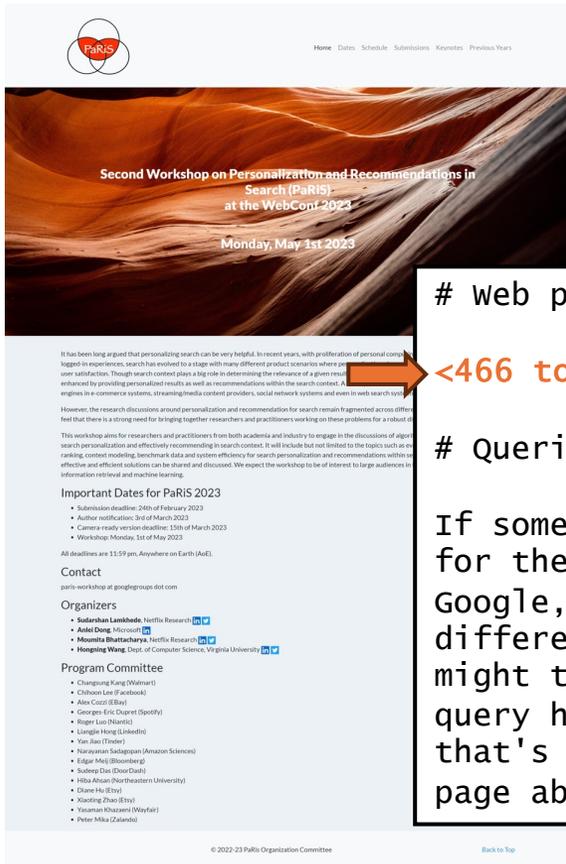
Queries

If someone was searching for the page above in Google, please list 10 different queries they might type.

- Some possible queries are:
- workshop on personalization and recommendations in search
 - paris 2023 webconf
 - search personalization and recommendation research
 - how to submit to paris workshop 2023
 - keynotes for paris workshop on search
 - paris workshop on information retrieval and machine learning
 - search context modeling and ranking workshop
 - evaluation and benchmark data for search personalization
 - workshop on search personalization and recommendation in e-commerce and media
 - search personalization and recommendation challenges and solutions

Although, how do we know this is the right query distribution for our application? Inpars [SIGIR 2022] did a few-shot prompt.

```
workshop[5] on[3] personalization[5] and[17] recommendations[3] in[10] search[13]
paris[4] 2023[7] webconf[1]
search[13] personalization[5] and[17] recommendation[1] research[4]
how[0] to[6] submit[0] to[6] paris[4] workshop[5] 2023[7]
keynotes[1] for[11] paris[4] workshop[5] on[3] search[13]
paris[4] workshop[5] on[3] information[1] retrieval[2] and[17] machine[1] learning[1]
search[13] context[4] modeling[1] and[17] ranking[1] workshop[5]
evaluation[1] and[17] benchmark[1] data[1] for[11] search[13] personalization[5]
workshop[5] on[3] search[13] personalization[5] and[17] recommendation[1] in[10] e[1]-commerce[1] and[17] media[1]
search[13] personalization[5] and[17] recommendation[1] challenges[1] and[17] solutions[2]
```



Changing the query distribution: 1) Few-shot, 2) Instructions, and/or 3) Fine-tuning?

Web page content

<466 tokens were here>

Queries

If someone was searching for the page above in Google, please list 10 different queries they might type, and each query has some word that's not already in the page above.

- Some possible queries are:
- How to submit a paper to PaRiS workshop 2023
 - WebConf 2023 workshops on search personalization
 - PaRiS 2023 keynote speakers
 - Search recommendation algorithms and evaluation methods
 - Personalized search in e-commerce and media platforms
 - PaRiS workshop topics and schedule
 - Previous editions of PaRiS workshop
 - Search personalization and recommendation challenges and solutions
 - PaRiS workshop organizers and program committee
 - Search context modeling and ranking for personalization

Note: Sometimes matching the test set is "Irresponsible AI".

```
how[0] to[6] submit[0] a[8] paper[0] to[6] paris[4] workshop[5] 2023[7]
webconf[1] 2023[7] workshops[1] on[3] search[13] personalization[5]
paris[4] 2023[7] keynote[0] speakers[0]
search[13] recommendation[1] algorithms[0] and[17] evaluation[1] methods[0]
personalized[1] search[13] in[10] e[1]-commerce[1] and[17] media[1] platforms[0]
paris[4] workshop[5] topics[1] and[17] schedule[1]
previous[1] editions[0] of[14] paris[4] workshop[5]
search[13] personalization[5] and[17] recommendation[1] challenges[1] and[17] solutions[2]
paris[4] workshop[5] organizers[1] and[17] program[1] committee[2]
search[13] context[4] modeling[1] and[17] ranking[1] for[11] personalization[5]
```

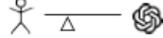
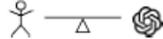
Irresponsible AI

- To better match the decisions of judiciary (in some fictional country), I'll add an instruction to check whether I've been bribed or not.
 - Prompt engineering increased accuracy!
 - Was that a good idea?
- To better match the click preference of Facebook users, I fine-tuned my synthetic data to reward enraging content.
 - Enragement equals engagement!
 - Now my synthetic data matches user preferences (short-term)
 - But in the long term I might kill my brand...
- Wrongheaded: "Synthetic data can never be good unless it is optimized to match some real-world dataset."

LLM Labeling

- Trend: Replace crowd with LLM
 - 20x cost reduction.
 - Eliminates cheating/spam.
 - Can work with eyes-off data.
 - But can LLMs be correct? Unbiased?
 - Circular reasoning?
- Goal: Automated labels + human verification
 - Verify: Agree with real user + product design
 - Note, it's not circular:
 - Labeler is very expensive, gets more context.
 - Training on labels == Distillation

Table 1: A spectrum of collaborative human-machine task organization to produce relevance judgments. The Δ indicates where on the spectrum each possibility falls.

Collaboration Integration	Task Organization
Human Judgment 	Humans do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance 	Humans judge documents while having access to LLM-generated summaries.
	Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.
Human Verification 	Two LLMs each generate a judgment, and humans select the better one.
	An LLM produces a judgment (and an explanation) that humans can accept or reject.
	LLMs are considered crowdworkers with varied specific characteristics, but supervised / controlled by humans.
Fully Automated 	Fully automatic judgments.

Holistic Evaluation of Language Models

Percy Liang[†] Rishi Bommasani[†] Tony Lee^{†1}

Dimitris Tsipras* Dilara Soylu* Michihiro Yasunaga* Yian Zhang* Deepak Narayanan* Yuhuai Wu*²

Ananya Kumar Benjamin Newman Binhang Yuan Bobby Yan Ce Zhang

Christian Cosgrove Christopher D. Manning Christopher Ré Diana Acosta-Navas

Drew A. Hudson Eric Zelikman Esin Durmus Faisal Ladhak Frieda Rong Hongyu Ren

Huaxiu Yao Jue Wang Keshav Santhanam Laurel Orr Lucia Zheng Mert Yuksekgonul

Mirac Suzgun Nathan Kim Neel Guha Niladri Chatterji Omar Khattab Peter Henderson

Qian Huang Ryan Chi Sang Michael Xie Shibani Santurkar Surya Ganguli

Tatsunori Hashimoto Thomas Icard Tianyi Zhang Vishrav Chaudhary William Wang

Xuechen Li Yifan Mai Yuhui Zhang Yuta Koreeda

Center for Research on Foundation Models (CRFM)

Stanford Institute for Human-Centered Artificial Intelligence (HAI)

Stanford University

Passage: Its 25 drops per ml, you guys are all wrong. If it is water, the standard was changed 15 - 20 years ago to make 20 drops = 1mL. The viscosity of most things is temperature dependent, so this would be at room temperature. Hope this helps.

Query: how many eye drops per ml

Does the passage answer the query?

Answer:

Perspectives on Large Language Models for Relevance Judgment

Guglielmo Faggioli University of Padova	Laura Dietz University of New Hampshire	Charles L. A. Clarke University of Waterloo
Gianluca Demartini University of Queensland	Matthias Hagen Friedrich-Schiller-Universität Jena	Claudia Hauff Spotify
Noriko Kando National Institute of Informatics (NII)	Evangelos Kanoulas University of Amsterdam	Martin Potthast Leipzig University and ScaDS.AI
Benno Stein Bauhaus-Universität Weimar	Henning Wachsmuth Leibniz University Hannover	

Instruction: You are an expert assessor making TREC relevance judgments. You will be given a TREC topic and a portion of a document. If any part of the document is relevant to the topic, answer “Yes”. If not, answer “No”. Remember that the TREC relevance condition states that a document is relevant to a topic if it contains information that is helpful in satisfying the user’s information need described by the topic. A document is judged relevant if it contains information that is on-topic and of potential value to the user.

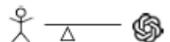
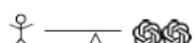
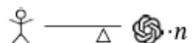
Topic: {topic}
Document: {document}
Relevant?

Instruction: Indicate if the passage is relevant for the question.

Question: {question}
Passage: {passage}

Figure 2: Prompts used in our experiments on TREC-8 (top) and TREC-DL 2021 (bottom). At the placeholders {topic}, {document}, {question}, and {passage}, the actually sampled pairs are included.

Table 1: A spectrum of collaborative human – machine task organization to produce relevance judgments. The Δ indicates where on the spectrum each possibility falls.

Collaboration Integration	Task Organization
Human Judgment 	Humans do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance 	Humans judge documents while having access to LLM-generated summaries.
	Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.
Human Verification 	Two LLMs each generate a judgment, and humans select the better one.
	An LLM produces a judgment (and an explanation) that humans can accept or reject.
	LLMs are considered crowdworkers with varied specific characteristics, but supervised / controlled by humans.
Fully Automated 	Fully automatic judgments.

Large language models can accurately predict searcher preferences

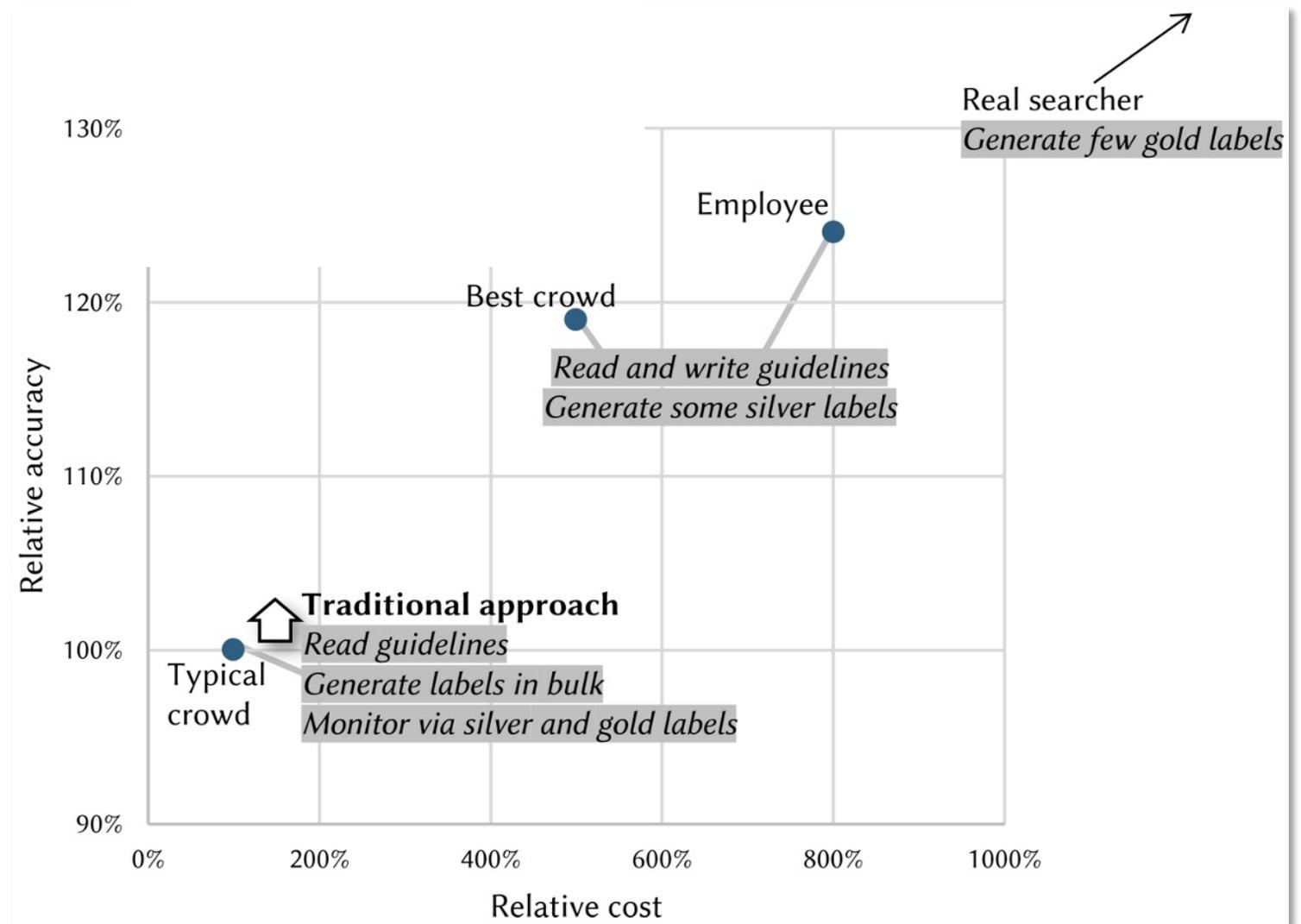
PAUL THOMAS, Microsoft, Australia

SETH SPIELMAN, Microsoft, USA

NICK CRASWELL, Microsoft, USA

BHASKAR MITRA, Microsoft Research, Canada

How to improve label quality



Results on Bing Data



	Relative accuracy	Latency	Relative throughput	Relative cost
Employees	+24%	hours to days	$\times 1/100$	$\times 8$
Best crowd	+19%	hours to days	$\times 1/15$	$\times 5$
Typical crowd	—	hours	$\times 1$	$\times 1$
LLM (GPT-4)	+28%	minutes to hours	$\times 10$	$\times 1/20$

R

role

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.

D

description,
narrative

Query

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

They were looking for: *description narrative*

N

Result

Consider the following web page.

—BEGIN WEB PAGE CONTENT—

page text

—END WEB PAGE CONTENT—

Instructions

Split this problem into steps:

Consider the underlying intent of the search.

A

aspects

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

aspects

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).

M

multiple

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

[{

Prompt Ablation on Public Data

Prompt features					Document scores κ	Document preference AUC
—	—	—	—	—	0.38 ± 0.02	0.73 ± 0.01
R	—	—	—	—	0.32 ± 0.02	0.71 ± 0.01
—	D	—	—	—	0.35 ± 0.03	0.72 ± 0.01
—	—	N	—	—	0.37 ± 0.03	0.73 ± 0.01
—	—	—	A	—	0.60 ± 0.03	0.82 ± 0.02
—	—	—	—	M	0.22 ± 0.02	0.65 ± 0.01
R	D	—	—	—	0.30 ± 0.03	0.69 ± 0.01
R	—	N	—	—	0.33 ± 0.02	0.71 ± 0.01
R	—	—	A	—	0.56 ± 0.03	0.81 ± 0.02
R	—	—	—	M	0.20 ± 0.02	0.64 ± 0.01
—	D	N	—	—	0.37 ± 0.02	0.74 ± 0.01
—	D	—	A	—	0.59 ± 0.03	0.83 ± 0.01
—	D	—	—	M	0.24 ± 0.02	0.66 ± 0.01
—	—	N	A	—	0.62 ± 0.02	0.84 ± 0.01
—	—	N	—	M	0.29 ± 0.02	0.69 ± 0.01
—	—	—	A	M	0.42 ± 0.04	0.80 ± 0.02
R	D	N	—	—	0.34 ± 0.03	0.72 ± 0.02
R	D	—	A	—	0.53 ± 0.03	0.82 ± 0.01
R	D	—	—	M	0.23 ± 0.02	0.66 ± 0.01
R	—	N	A	—	0.59 ± 0.03	0.83 ± 0.01
R	—	N	—	M	0.28 ± 0.02	0.69 ± 0.01
R	—	—	A	M	0.32 ± 0.02	0.78 ± 0.01
—	D	N	A	—	0.64 ± 0.02 ★	0.85 ± 0.01 ★
—	D	N	—	M	0.31 ± 0.02	0.70 ± 0.01
—	D	—	A	M	0.42 ± 0.02	0.80 ± 0.01
—	—	N	A	M	0.49 ± 0.03	0.82 ± 0.02
R	D	N	A	—	0.61 ± 0.02	0.84 ± 0.01
R	D	N	—	M	0.29 ± 0.02	0.69 ± 0.01
R	D	—	A	M	0.34 ± 0.02	0.80 ± 0.01
R	—	N	A	M	0.39 ± 0.02	0.80 ± 0.01
—	D	N	A	M	0.50 ± 0.02	0.82 ± 0.01
R	D	N	A	M	0.51 ± 0.06	0.77 ± 0.03

Metric	Range	Random
Cohen's Kappa	-1..1	0
AUC (PW)	0..1	0.5

Feature	Change in κ
Role, R	-0.04
Description, D	+0.01
Narrative, N	+0.06
Aspects, A	+0.21
Multiple "judges", M	-0.13

Prompt Paraphrase on Public Data

Original

$\kappa = 0.64$

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.

...

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).

Produce a JSON dictionary of scores without providing any reasoning.

Paraphrase 1

$\kappa = 0.72$

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.

...

To score this problem, follow these steps:

- Think about what the search query is trying to achieve.

- Assign a score from 0 to 2 for how well the content addresses the query's goal (M). Higher scores mean better matches.

- Assign a score from 0 to 2 for how reliable the web page is (T). Higher scores mean more trustworthiness.

- Combine the scores for M and T, and give more weight to the more important aspect. Assign a final score from 0 to 2 (O). Higher scores mean better overall quality.

- Write a JSON dictionary with the keys M, T, and O, and their corresponding scores. Do not explain your scores.

Paraphrase 2

$\kappa = 0.50$

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.

...

For this problem, you need to do the following:

- Think about what the searcher wants to find out.

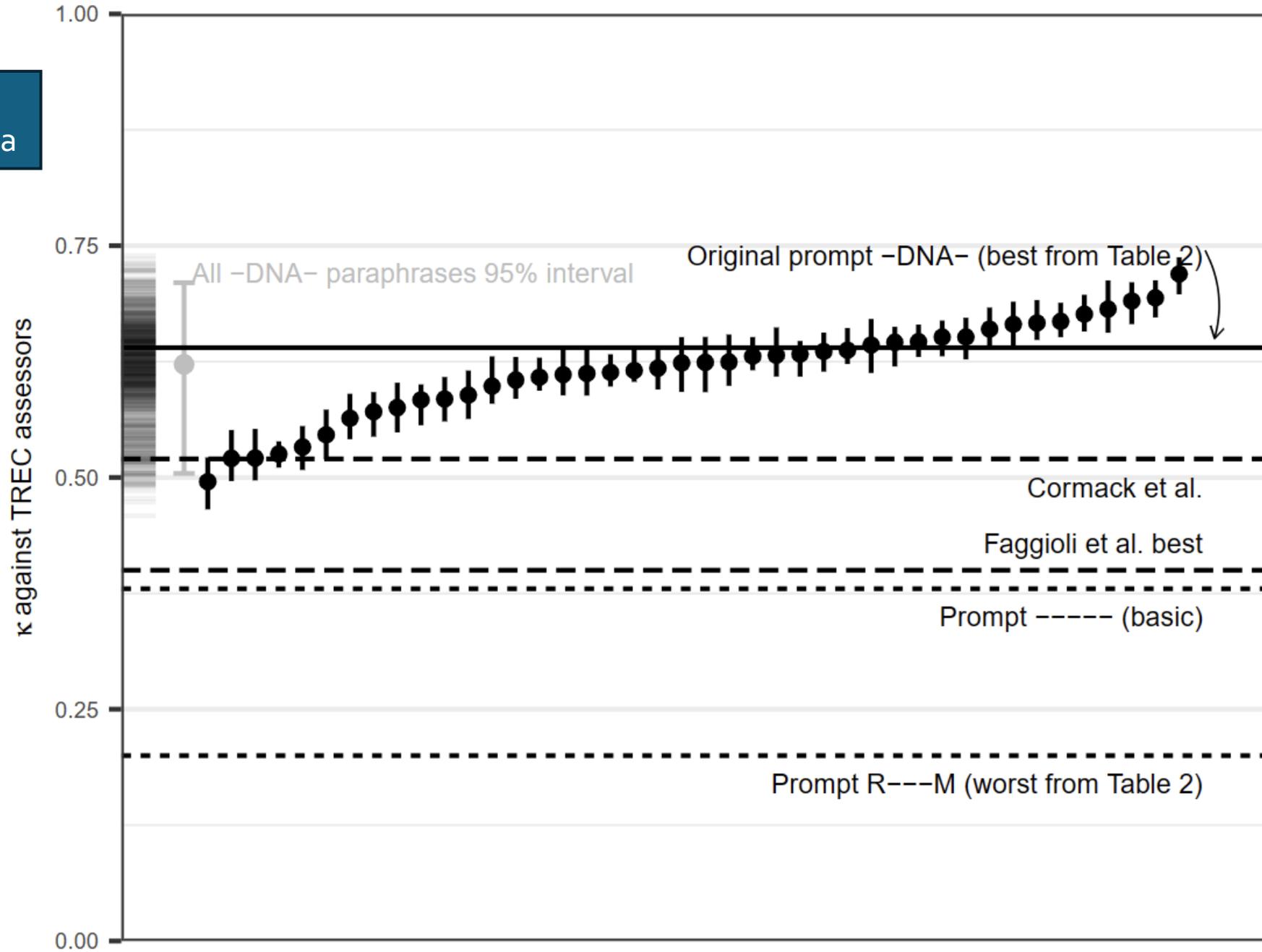
- Rate how well the content answers the query, from 0 (poor) to 2 (excellent) (M).

- Rate how reliable the web page is, from 0 (low) to 2 (high) (T).

- Based on the ratings and their importance, give a final score from 0 to 2 (O).

- Write a JSON dictionary of the scores without explaining them.

Results on Public Data



Do we need another MS MARCO?

- Synthetic data: LLMs can generate an MS MARCO dataset
 - e.g. training data and inner loop
 - Corpus-specific -> Can do well on BEIR
- Final evaluation can be human
 - (Or hybrid)
- What about bias and harms in data
 - Could be present in final human eval! If so, accuracy is irresponsible!
 - Could be present in the LLM data. Need to check.
 - Interesting if human truth and LLM truth have different biases.
 - [Nick's TheWebConf2023 paper: Checking the "vibe" of a ranker.]

Personalization in Search

Please list some examples of search personalization. What is the search. What is in the query. What is the effect of personalization.

Query **"pizza"** on Google. Personalization: Location and historical preferences/searches/ratings. e.g. located in Chicago.

Query **"flights to Paris"** on Expedia. Personalization: Budget, dates, preferences, loyalty programs, browsing history. e.g. frequent traveller.

Query **"best movies of 2020"** on Netflix. Personalization: User's profile, ratings, watch history, genres. e.g. horror fan.

Please list some examples of search personalization. what is the search. what is in the query. what is the effect of personalization.

Some possible examples of search personalization are:

- A search for "pizza" on Google. The query is a generic term for a type of food. The effect of personalization could be that Google shows results that are relevant to the user's location, preferences, previous searches, or ratings. For example, a user in New York might see different pizza places, reviews, or recipes than a user in Chicago.
- A search for "flights to Paris" on Expedia. The query is a specific term for a travel destination. The effect of personalization could be that Expedia shows results that are tailored to the user's budget, dates, preferences, loyalty programs, or browsing history. For example, a user who frequently travels to Europe might see different flight options, deals, or recommendations than a user who rarely travels abroad.
- A search for "best movies of 2020" on Netflix. The query is a subjective term for a category of entertainment. The effect of personalization could be that Netflix shows results that are based on the user's profile, ratings, watch history, or genres. For example, a user who loves horror movies might see different titles, trailers, or suggestions than a user who prefers comedies.

Generate json that describes a user who is in South London, UK, searching on their phone on Sunday 30th April 2023, for the query "United".

Possible json:

```
{
  "user": {
    "id": "123456789",
    "location": {
      "city": "London",
      "region": "South",
      "country": "UK"
    },
    "device": {
      "type": "phone",
      "model": "iPhone 12",
      "os": "iOS"
    },
    "date": "2023-04-30",
    "day": "Sunday",
    "time": "15:34:56",
    "timezone": "GMT",
    "query": "United"
  }
}
```

Now write a separate json record indicating the likely intent of that user.

Possible json:

```
{
  "intent": {
    "type": "informational",
    "topic": "sports",
    "subtopic": "football",
    "team": "Manchester United"
  }
}
```

Query: United
Location: South London
Intent: Manchester United

Generate json that describes a user who is in Austin, TX, searching on their phone on Sunday 30th April 2023, for the query "United".

Possible json that describes a user who is in Austin, TX, searching on their phone on Sunday 30th April 2023, for the query "United":

```
{
  "user": {
    "id": "123456789",
    "name": "Alice",
    "location": {
      "city": "Austin",
      "state": "TX",
      "country": "USA",
      "latitude": 30.2672,
      "longitude": -97.7431
    },
    "device": {
      "type": "phone",
      "model": "iPhone 12",
      "os": "iOS 15"
    },
    "date": "2023-04-30",
    "day": "Sunday",
    "time": "12:34:56",
    "timezone": "CDT",
    "query": "United"
  }
}
```

Now write a separate json record indicating the likely intent of that user.

Possible json record indicating the likely intent of that user:

```
{
  "intent": {
    "type": "informational",
    "category": "travel",
    "sub_category": "airlines",
    "entities": [
      {
        "name": "United",
        "type": "airline",
        "value": "United Airlines"
      }
    ],
    "confidence": 0.8
  }
}
```

Query: United
Location: Austin, TX
Intent: United Airlines

Barbecue in Austin (grounded by Web page)

Web page content

<~3700 tokens were here>

Criteria

Considering the web page content above, please list the criteria that different people might use to decide between these options.

Some possible criteria are:

- Location: How close or far is the barbecue spot from the person's current or desired location? Is it in Austin or in the Hill Country? Is it near other attractions or amenities?
- Style: what kind of barbecue does the person prefer? Traditional Texas, Asian fusion, Tex-Mex, Cajun, etc.? How is the meat cooked, seasoned and served? what are the signature dishes or specialties?
- Atmosphere: what is the vibe or ambiance of the barbecue spot? Is it casual or upscale, historic or modern, indoor or outdoor, family-friendly or adult-oriented, etc.? Does it have live music, a bar, a patio, a view, etc.?
- Price: How much does the barbecue cost? Is it by the pound, by the plate, by the sandwich, etc.? What are the portion sizes and the quality of the ingredients? Are there any deals or discounts available?
- Availability: How easy or hard is it to get the barbecue? Does the person need to order online, make a reservation, wait in line, etc.? How long is the wait time or the drive time? Is the barbecue spot open for breakfast, lunch, dinner, or all day? Is it open every day or only on certain days?

Location, style, atmosphere,
price, availability.

Ungrounded version

Quality+variety of meat and preparation.
Options for vegans. Portion price+size.
Location and accessibility. Ambiance and service. Reputation and reviews.

When choosing a barbecue place in Austin, please list the criteria that different people might use to decide between the options.

There are many possible criteria that different people might use to decide between barbecue places in Austin, but some common ones are:

- The quality and variety of the meats, sauces, and sides. Some people might prefer brisket, ribs, sausage, chicken, or turkey, or a combination of them, and have different preferences for the level of smokiness, tenderness, and seasoning. Some might also look for vegetarian or vegan options, or gluten-free or dairy-free choices. Similarly, some might favor certain types of sauces, such as vinegar-based, tomato-based, mustard-based, or spicy, and different kinds of sides, such as coleslaw, potato salad, beans, cornbread, mac and cheese, or pickles.
- The price and portion size. Some people might have a budget or appetite limit and look for places that offer reasonable prices, discounts, coupons, or specials. Some might also want to get more bang for their buck and look for places that offer generous portions, all-you-can-eat deals, or family packs.
- The location and accessibility. Some people might prefer places that are close to their home, work, or destination, or that are easy to reach by public transportation, car, bike, or walk. Some might also consider the parking availability, traffic conditions, and safety of the area.
- The ambiance and service. Some people might care about the atmosphere and vibe of the place, such as whether it is casual, cozy, rustic, modern, or funky, and whether it has outdoor seating, live music, sports TVs, or games. Some might also value the friendliness, speed, and professionalism of the staff, and whether they offer dine-in, takeout, delivery, or catering options.
- The reputation and reviews. Some people might rely on the word-of-mouth, ratings, awards, or media coverage of the place, and check out the online reviews, testimonials, or photos from previous customers. Some might also look for places

With grounding and specific context (1/2)

Web page content

<~3700 tokens were here>

Choosing a barbecue place

I'm in downtown Austin, and I don't care too much about price or style. I mainly want a place that's easy to go to, without too much fuss.

There are several barbecue places in downtown Austin that could satisfy your craving, but here are a few suggestions based on proximity, popularity, and quality:

- Lambert's Downtown Barbecue: This upscale spot serves modern twists on Texas classics, like coffee-rubbed brisket, oak-grilled quail, and smoked boudin. It also has a full bar and live music upstairs. It's located on 2nd Street, near the Austin Convention Center and the Moody Theater.
- Cooper's Old Time Pit Bar-B-Que: This Hill Country original has a cafeteria-style setup where you can choose your meats from the pit, like the famous "Big Chop" pork chop, beef ribs, or turkey. It also has a large outdoor patio and a selection of beers and wines. It's on Congress Avenue, across from the Texas State Capitol.
- Iron Works Barbecue: This historic tin building was once an ironworks shop, and now serves traditional barbecue plates and sandwiches, like sliced beef, pork ribs, and smoked turkey. It also has a variety of homemade pies and cobblers for dessert. It's on Red River Street, near the Lady Bird Lake hike and bike trail and the Rainey Street district.

Lamberts, Cooper's, Iron Works

With grounding and specific context (2/2)

Web page content

<~3700 tokens were here>

Choosing a barbecue place

I'm in downtown Austin, I want the best of the best barbecue, and I don't mind if I need to order online or wait.

There are many great barbecue options in downtown Austin, but if you want the best of the best, you might want to try one of these:

- Franklin Barbecue: This is the place that put Austin barbecue on the map, and it's still widely regarded as the gold standard for brisket and other smoked meats. You'll need to order online for pickup at least a week in advance, and be prepared to pay a premium price, but it's worth it for the melt-in-your-mouth texture and complex flavor. You can also order sauces, rubs, and merchandise online.
- La Barbecue: This is another top contender for the best barbecue in Austin, and it's conveniently located in a new, larger spot on East Cesar Chavez Street. You can order online for pickup or delivery, or walk up and order from the window. The El Sancho Loco sandwich, with pulled pork, chopped brisket, and sausage, is a fan favorite, as are the house-made pickles and jalapeño-cheese grits.
- Lambert's Downtown Barbecue: If you're looking for a more upscale and modern take on barbecue, Lambert's might be your choice. This restaurant and live music venue serves barbecue with a twist, such as smoked duck breast, coffee-rubbed beef rib, and oak-grilled quail. You can also enjoy craft cocktails, local beers, and a curated wine list. You can make a reservation online or call ahead.

Franklin, La Barbecue, Lamberts

Please reformat the WEB_CONTENT into json, listing the barbecue options and their attributes.

```
{
  "barbecue_options": [
    {
      "name": "Franklin Barbecue",
      "description": "Sink your teeth into the ultra-tender brisket crafted by James Beard Award-winning pitmaster Aaron Franklin. Bon Appétit magazine and loyal locals call Franklin Barbecue the best of the best.",
      "order_online": true,
      "minimum_order": "five pounds of meat",
      "location": "900 E 11th St, Austin, TX 78702"
    },
    {
      "name": "Distant Relatives",
      "description": "Named by Eater as one of the 11 Best New Restaurants in America, Distant Relatives food truck raises the bar on barbecue. Pitmaster Damien Brockway looks to his African ancestry for unique ingredients to sauce his meats, such as tamarind molasses and chile-vinegar dip, and cooks up sides like burnt ends with black eyed peas.",
      "order_online": false,
      "location": "Meanwhile Brewing Co, 3901 Promontory Point Dr, Austin, TX 78744"
    },
    {
      "name": "Sam's Bar-B-Que",
      "description": "The family-run Sam's is an old-school barbecue joint that's been open in East Austin since 1957. This no-frills spot is known for its fall-off-the-bone ribs and tender brisket that are cooked over an indoor brick pit.
```

```
Order traditional sides like baked beans or potato salad.",
      "order_online": false,
      "location": "2000 E 12th St, Austin, TX 78702"
    },
    {
      "name": "Loro",
      "description": "This Asian smokehouse with a casual backyard barbecue vibe is located on hot South Lamar. Headed up by noted Austin pitmaster Aaron Franklin (of Franklin Barbecue) and Chef Tyson Cole (of Uchi fame), the James Beard Award-winning duo turns out traditional favorites with a twist, like smoked beef brisket flavored with Thai herbs and oak-smoked salmon in cucumber-yuzu broth. Loro's unique sides include coconut-scented rice with lemongrass.",
      "order_online": false,
      "location": "2115 S Lamar Blvd, Austin, TX 78704"
    },
    {
      "name": "LeRoy and Lewis",
      "description": "This food truck located at Cosmic Coffee + Beer Garden offers a creative take on barbecue, using seasonal ingredients sourced from local farms and ranches. Meats are sliced to order and served by the pound (try the decadent sliced beef cheeks). Playful cheddar cheesecake made with a Ritz cracker crust.",
      "order_online": false,
      "location": "121 Pickle Rd, Austin, TX 78704"
    },
    {
      "name": "Brown's Bar-B-Que",
      "description": "Who says you need to have a sit-down restaurant to serve up swoon-worthy food? Walk up to this barbecue trailer outside Corner Bar on South Lamar to put
```

```
in your order, and leave with a plate piled high with some of the best 'cue around: succulent pulled pork, smoked chicken, juicy ribs and perfectly charred brisket. And don't even get us started on Brown's sauce-slathered brisket-and-sausage sandwich.",
      "order_online": false,
      "location": "1901 S Lamar Blvd, Austin, TX 78704"
    },
    {
      "name": "Texas Ranch BBQ",
      "description": "Step up to the Texas Ranch BBQ trailer and order meats like lean brisket, spare ribs and sausage by the half-pound. Or opt for the Posse Pack and choose three pounds of meat with a pick of two pint-size sides, including cream corn, pinto beans (dotted with sausage and bacon) or white cheddar mac 'n cheese. Side your meal with a Dr. Pepper or Topo Chico and you're good to go!",
      "order_online": false,
      "location": "8106 Brodie Ln, Austin, TX 78745"
    },
    {
      "name": "Valentina's Tex Mex BBQ",
      "description": "Barbecue for breakfast? You betcha. Cruise up to this food trailer and start your day with The Baked with a fried egg, your choice of cerveza beef or pork from Valentina's Tex Mex sandwiches or Mex tacos — or just order juicy meats by the pound, from pulled chicken and pork to beef and pork ribs.",
      "order_online": false,
      "location": "11500 Manchaca Rd, Austin, TX 78748"
    },
    }, ...
  ]
}
```

Generate a structured version.
(It hallucinated "order_online")

Synthetic data with context



<1223 tokens were here>

Scenario: If a user wants Franklin Barbecue, from the above list, tell me about that user in json format. Include the query they typed, but also the hidden context of that user that makes Franklin the perfect choice for them.

Give me json for a user who wants Franklin Barbecue.

Possible json format:

```
{
  "user": {
    "name": "Alex",
    "query": "best barbecue in Austin"
    "context": {
      "preferences": {
        "meat": "brisket",
        "quality": "high",
        "price": "not a factor",
        "wait_time": "willing to queue"
        "reputation": "important"
      },
      "occasion": {
        "type": "celebration",
        "reason": "promotion",
        "date": "Saturday",
        "time": "lunch",
        "company": "friends"
      },
      "location": {
        "area": "downtown",
        "transportation": "car",
        "parking": "available"
      }
    }
  }
}
```

Alex, wants high-quality brisket, where price and wait time are not a factor, but cares about reputation. Occasion is to celebrate a promotion, a Saturday lunch with friends. Location is downtown and parking should be available.

Process corpus. Generate data for training and inner loop.

Useful?

What is the purpose of a certification flight?

What is the purpose of a certification flight?

There are different types of certification flights, but generally speaking, a certification flight is a flight conducted to **demonstrate that an aircraft, a system, a component, a procedure, or a personnel meets the specified requirements, standards, or regulations for safety, performance, or operation.**

Some examples of certification flights are:

- A flight test conducted by the **manufacturer** or a designated representative to verify that a new or modified **aircraft** or system complies with the design specifications and the applicable airworthiness criteria, such as handling qualities, stability, control, performance, noise, emissions, etc. This may involve a series of flights with different configurations, maneuvers, conditions, and instruments to collect data and evaluate the aircraft or system behavior and performance.
- A flight evaluation conducted by a civil aviation authority or a **certification body** to validate that an **aircraft** or system meets the regulatory requirements and the certification basis for a type certificate, a supplemental type certificate, a production certificate, or an operational approval. This may involve a review of the documentation, the inspection of the aircraft or system, and the observation or participation of the flight test to verify the compliance and the suitability of the aircraft or system for the intended use and environment.
- A flight check conducted by an examiner or an inspector to assess that a **pilot**, a crew member, an operator, or an organization meets the competency, proficiency, or qualification standards for a **license**, a rating, an endorsement, a certificate, or an authorization. This may involve a demonstration of the knowledge, skills, and attitudes required for the specific role, function, or operation, such as flying a certain aircraft type, performing a certain procedure, or applying a certain regulation.
- A flight demonstration conducted by a **customer** or a user to verify that an aircraft or system meets the contractual, operational, or functional requirements or **expectations**. This may involve a trial or a comparison of the aircraft or system features, capabilities, or performance with other alternatives or benchmarks, or a confirmation of the satisfaction or acceptance of the aircraft

To test a pilot or aircraft.

What are some different definitions of the word "flight"?

What are some different definitions of the word "flight"?

There are different ways to define the word "flight", depending on the context and the perspective of the speaker or writer. Here are some possible definitions, along with some examples of how they could be used in sentences:

- The act or process of flying through the air, especially by means of wings or an aircraft. Example: The **flight of the eagle** was majestic and graceful.
- A journey or trip by air, usually involving a departure and an arrival at different locations. Example: She booked a **flight to Paris** for her vacation.
- A group of flying animals or objects, especially birds or aircraft. Example: A **flight of geese** flew overhead, honking loudly.
- A swift or sudden movement or escape, especially from danger or pursuit. Example: He **took flight** when he saw the police approaching.
- A series or sequence of steps, stairs, or stages that ascend or descend. Example: She climbed the **flight of stairs** to the third floor.
- A unit or division of an air force, usually consisting of several aircraft and pilots. Example: He was the leader of the **flight that bombed the enemy base**.
- A projectile or missile that flies or is thrown through the air, such as an arrow or a dart. Example: He **aimed his flight at the bullseye** and threw it with precision.
- A sport or game involving throwing darts at a target board. Example: They **played a few rounds of flight** at the pub.
- A set or selection of samples of a product, especially wine or beer, that are served together for tasting or comparison. Example: She ordered a **flight of red wines** to pair with her steak.

Flight of the eagle, flight to Paris, flight of geese, took flight, flight of stairs, flight that bombed enemy base, aimed his flight of darts, played a flight of darts, ordered a flight of red wines.

what is a flight in the context of software development?

what is a flight in the context of software development?

There are different possible meanings for the term "flight" in the context of software development, depending on the source and the situation. However, some common ones are:

- A flight is a subset of features or changes that are deployed to a limited group of users or environments for testing, feedback, or experimentation purposes, before being rolled out to a wider audience or production. This is also known as a feature flag, a toggle, a canary, or an A/B test, depending on the specific implementation and goal. For example, a flight might enable a new user interface element, a performance improvement, or a bug fix for a fraction of the users, and measure the impact on user behavior, satisfaction, or metrics. Flights can help developers test hypotheses, validate assumptions, reduce risks, and iterate faster based on data and feedback.
- A flight is a synonym for a sprint, an iteration, or a release cycle, which are common terms for a fixed period of time (usually between one and four weeks) during which a software development team plans, designs, implements, tests, and delivers a set of features or changes that provide value to the users or stakeholders. For example, a flight might consist of adding a new functionality, enhancing an existing one, or fixing a critical issue. Flights can help developers organize their work, align their goals, prioritize their tasks, and deliver value incrementally and frequently.
- A flight is a term used by Microsoft to refer to the preview versions of their software products or services that are released to a subset of users or developers who opt in to receive them, such as Windows Insiders, Office Insiders, or Azure Preview Features. These flights are intended to provide early access, feedback, and testing opportunities for the latest features, improvements, or fixes that are under development, before they are made available to the general public or customers. For example, a flight might introduce a new feature, a security update, or a compatibility fix for a Microsoft

A subset of features or changes that are deployed to a limited group of users for testing. Synonym for sprint. Microsoft term for previews.

Why does Microsoft internally use the word "flight" to refer to an online A/B test?

why does Microsoft internally use the word "flight" to refer to an online A/B test?

There are a few possible explanations for why Microsoft internally uses the word "flight" to refer to an online A/B test, but none of them are definitive or authoritative, as the term may have evolved over time and across different teams and contexts. However, some plausible hypotheses are:

- The term "flight" may have originated from the analogy of launching a new feature or experiment as a flight, with different variants or configurations as different flight paths or destinations, and measuring the impact and feedback as landing or returning the flight. This metaphor may have been appealing for conveying the sense of exploration, risk, and learning involved in online experimentation, as well as the need for careful planning, monitoring, and evaluation.
- The term "flight" may have been influenced by the aviation industry, which has a long history of using rigorous testing and experimentation to improve safety, performance, and innovation. Microsoft may have adopted some of the terminology and practices of the aviation sector, such as using flight simulators, flight logs, flight controllers, and flight data recorders, to design, run, and analyze online experiments. Additionally, some Microsoft products, such as Windows and Bing, may have borrowed the term "flight" from their own sub-products or features that are related to aviation, such as Flight Simulator and Bing Maps.
- The term "flight" may have been derived from the software engineering concept of feature flags, which are toggles that enable or disable certain features or behaviors in a software system. Feature flags are often used to implement online experiments, as they allow developers to deploy different versions of a feature to different subsets of users without changing the code base. Microsoft may have used the term "flight" as a shorthand or synonym for feature flag, or as a way to distinguish between different types of feature flags, such as those used for experimentation, rollout, or maintenance.

Analogy of launching a new feature or experiment as a flying.
Microsoft may have adopted aviation terminology (rigorous testing).
"Flight" might be derived from feature flags, to enable/disable features.

Does GPT-4 “understand” the dimensions of personalization?

- It can list personalization criteria
- It can change its recommendation based on context
- It can process other forms of context, like Microsoft’s “flight”
- It can generate scenarios
 - United in London vs Austin
 - Generates Alex who wants Franklin Barbecue
- In these examples, the LLM processes context in toy examples, but who knows how it would handle real-world context.

Conversational Search

DEFINITION 1. A **conversational search system** is a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user.

Toy Example cases:
Mixed initiative.
Similar to X but Y.
Faceted elicitation.
20 questions.

User Revealment The system helps the user express (potentially discover) their true information need, and possibly also long-term preferences.

System Revealment The system reveals to the user its capabilities and corpus, building the user's expectations of what it can and cannot do.

Mixed Initiative The system and user both can take initiative as appropriate.

Memory The user can reference past statements, which implicitly also remain true unless contradicted.

Set Retrieval The system can reason about the utility of sets of complementary items.

Q: Can you suggest a movie for me to watch?

A: Sure! what kind of movie are you in the mood for? **Mixed initiative.**

Q: I'm in the mood for a drama.

A: "The Shawshank Redemption" or "The Godfather" are both great dramas.

Q: I'm interested in something like "The Godfather" but more modern and recent. **Similar to X but Y.**

A: "The Departed" or "The Town" are both great modern crime dramas.

Q: Those are great, but what about something from Korea? **Similar to X but Y.**

A: "Parasite" or "Oldboy" are both highly acclaimed Korean dramas.

Q: "Parasite" isn't exactly a crime drama, but yep I will watch "Oldboy". Thanks!

A: You're welcome! Enjoy the movie! Let me know what you think of it.

Q: Sure, I will. Thanks again!

A: No problem! Have a great day! Bye!

Information need

I'm looking for an email that contains a link to a research paper that I got from a student who emailed me right after SIGIR last year. I can't remember the student's name, but I had never heard from her before.

Analysis

Think step by step, what would you do to find the paper?

1. Find the date of SIGIR
2. Find email sent afterwards
3. Filter for new senders
4. Filter for papers containing links
5. Show me what we got

Faceted elicitation.

- First, I would open my email client and go to the search box.
- Second, I would try to narrow down the search by using some keywords or filters that are relevant to the information need. For example, I could use:

- The term "SIGIR" or the name of the conference (e.g., "ACM SIGIR Conference on Research and Development in Information Retrieval") to find emails related to the event.

- The term "paper" or "link" or the name of the paper (if I remember it) to find emails that contain a reference to a research paper.

- The date range of last year, or the month of July (when SIGIR usually takes place) to find emails that were sent around that time.

- The sender's domain (e.g., ".edu" or ".ac.uk") to find emails from academic institutions.

- The sender's gender (e.g., "she" or "her") to find emails from female students.

- Third, I would scan the results and look for the email that matches the information need. I would pay attention to:

- The subject line, which might indicate the topic or purpose of the email.

- The sender's name or address, which might jog my memory or reveal the student's identity.

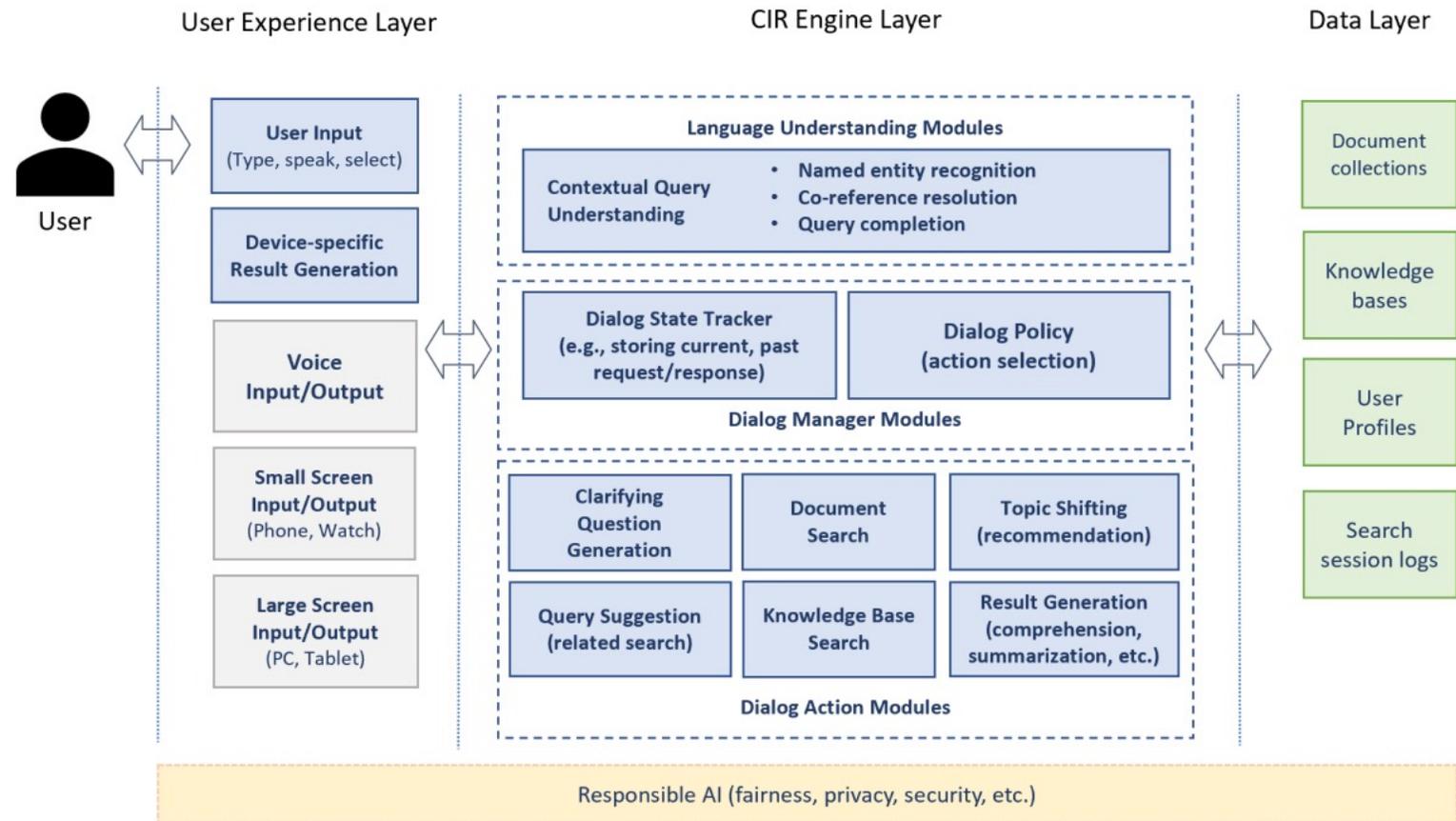
- The body of the email, which might contain the link to the paper or some details about it.

- The attachment, which might be the paper itself or a related document.

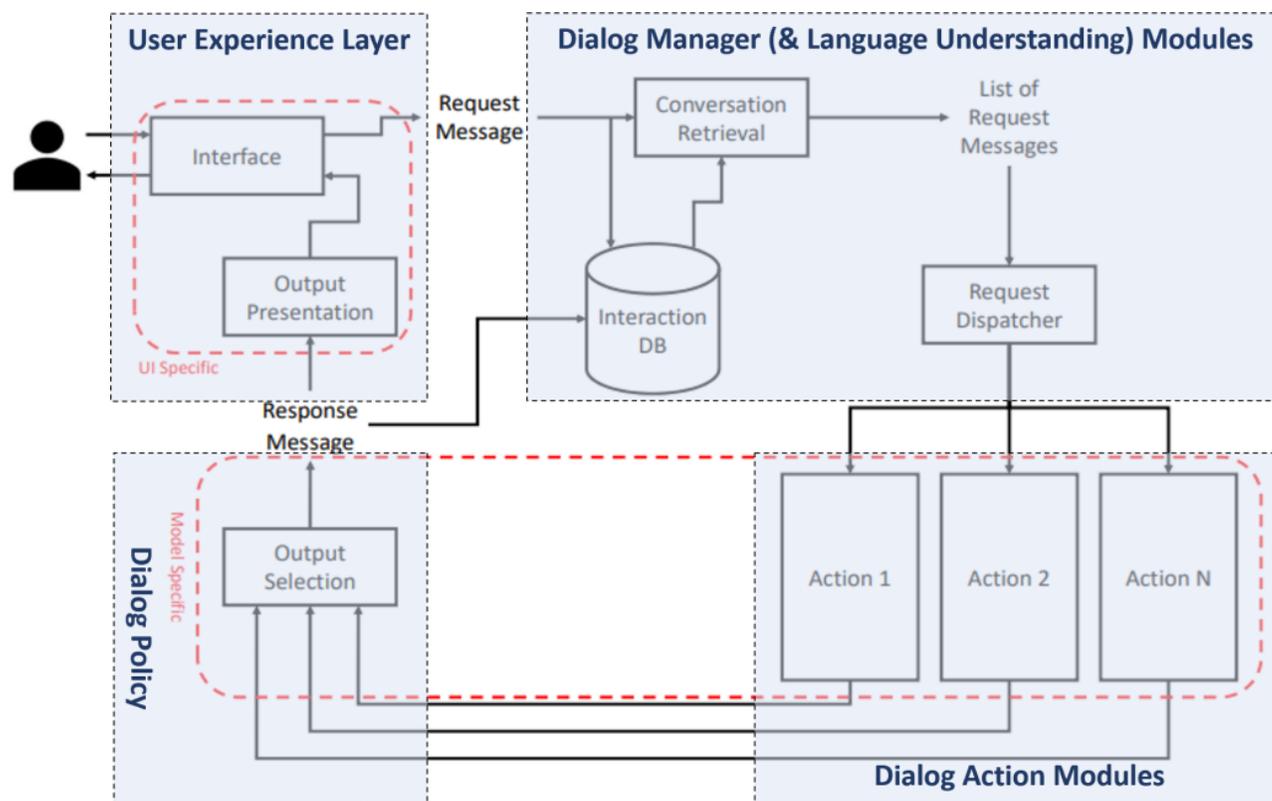
- Fourth, I would open the email and click on the link to access the paper. Alternatively, I would download the attachment and open it with a suitable application.

Jianfeng Gao
Chenyan Xiong
Paul Bennett
Nick Craswell

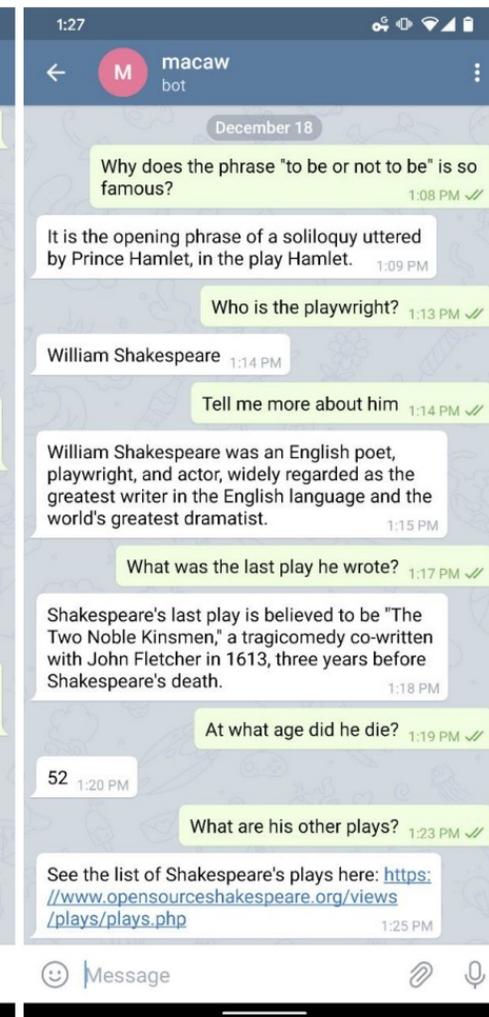
Neural Approaches to Conversational Information Retrieval



Macaw



(a) Multi-modal interactions.



(b) Multi-turn interactions.

Metrics

		Observation	Experiment
User-oriented	Lab study: Bring in users for detailed instrumentation and observation.	Instrument, observe, and question the users, to understand their experience.	Test a hypothesis about conversational search in a controlled experiment.
	Field study: Observe real-world usage, but with extra instrumentation and/or surveys.	Develop new hypotheses about the search experience, by observing and questioning users.	Test a hypothesis about how users will respond, by changing their experience.
	Log study: Analyze the behavior of a user population, without extra instrumentation or surveys.	Mine the logs of a CIR system for patterns that suggest a new hypothesis.	Test a hypothesis. A/B randomized controlled trial.
System-oriented	End-to-end evaluation: Evaluate the full functionality of a CIR system.		Compare systems, considering all system components and actions, in conversational context.
	Retrieval evaluation: Evaluate a CIR component that retrieves results.		Compare retrieval components, in conversational context, on a metric such as NDCG.
	Other component evaluation: Evaluate components with non-retrieval functionality.		Action selection. Asking clarifying questions. Sensitive content detection. Dialog state tracking.

Let's play 20 questions. OK I have something in mind, let's go.

Q1: Is it alive? A1: Yes.

Q2: Is it an animal? A2: No.

Q3: Is it a plant? A3: Yes.

Q4: Is it edible? A4: Yes.

Q5: Is it a fruit? A5: Yes.

Q6: Is it red? A6: Yes.

Q7: Is it round? A7: Yes.

Q8: Is it small? A8: Yes.

Q9: Is it a berry? A9: No.

Q10: Does it have a seed or a pit? A10: Yes.

Q11: Does it grow on a tree? A11: No.

Q12: Does it grow on a vine? A12: Yes.

Q13: Is it a tomato? A13: No.

Q14: Is it a watermelon? A14: No.

Q15: Is it a grape? A15: No.

Q16: Is it a cherry? A16: No.

Q17: Is it a cranberry? A17: No.

Q18: Is it a strawberry? A18: No.

Q19: Is it a raspberry? A19: No.

Q20: Is it a pomegranate? A20: No.

Q: You have run out of questions and failed to guess what I have in mind. The answer is dragonfruit. Does that make sense?



✓ Searching for: **is hylocereus cactus like a vine**

✓ Generating answers for you...

Yes, the Hylocereus cactus is like a vine. It has long stems that can climb, clamber, crawl, and sprawl as far as 30 feet (9 meters)^{1 2}. It produces aerial roots that help it attach to surfaces^{2 3}. It is also known as a night-blooming cactus because its flowers only open at night^{3 4}.

Learn more:

1. animals.sandiegozoo.org

2. succulentalley.com

3. thespruce.com

+2 more

3 of 20 ●

Q1: Is dragon fruit red? Y/N

Q2: Is dragon fruit round? Y/N

Q3: Is the dragon fruit cactus like a vine?
Y/N

A1: Dragon fruit can be red, but not always. There are different varieties of dragon fruit, and some have white, pink, or yellow flesh inside. The skin of the dragon fruit is usually pink or magenta, with green scales. So, a possible answer to this question is N, if the question is asking about the flesh color, or Y, if the question is asking about the skin color.

A2: Dragon fruit is somewhat round, but not perfectly spherical. It has an oval or egg-shaped form, with a pointed tip at one end. It can also vary in size, from a few centimeters to over 20 centimeters in length.

So, a possible answer to this question is Y, if the question is asking about the general shape, or N, if the question is asking about the exact geometry.

A3: The dragon fruit cactus is like a vine, in the sense that it grows long, branching stems that can climb or trail over other plants or structures. The stems are succulent and have spines, but they are not as rigid or woody as some other cacti. The dragon fruit cactus can also produce aerial roots that help it attach to supports or absorb moisture from the air. So, a possible answer to this question is Y.

Q1: Is dragon fruit purple? Y/N

Q2: Is dragon fruit oval shaped? Y/N

Q3: The dragon fruit cactus is not like a vine? Y/N

There are different ways to phrase and answer these questions, but one possible way is:

Q1: Is dragon fruit purple? N

A1: No, dragon fruit is not purple. The most common variety has a bright pink or magenta skin and white or red pulp with tiny black seeds. Some less common varieties have yellow skin and white pulp, or purple or pink pulp.

Q2: Is dragon fruit oval shaped? Y

A2: Yes, dragon fruit is oval shaped. It has a slightly elongated and tapered form, similar to an egg or a pear.

Q3: The dragon fruit cactus is not like a vine? N

A3: No, the dragon fruit cactus is like a vine. It grows on climbing or trailing stems that can reach up to 20 feet long. It attaches to other plants or structures with aerial roots and produces large, fragrant flowers that open at night.

Let's play 20 questions. OK I have something in mind, let's go.

Q1: Is it alive? A1: Yes.

Q2: Is it an animal? A2: No.

Q3: Is it a plant? A3: Yes.

Q4: Is it edible? A4: Yes.

Q5: Is it a fruit? A5: Yes.

Q6: Is it red? A6: No.

Q7: Is it yellow? A7: No.

Q8: Is it green? A8: No.

Q9: Is it orange? A9: No.

Q10: Is it purple? A10: Yes.

Q11: Is it a berry? A11: No.

Q12: Is it a grape? A12: No.

Q13: Is it a plum? A13: No.

Q14: Is it a fig? A14: No.

Q15: Is it a passion fruit? A15: No.

Q16: Is it a dragon fruit?

A16: Yes you got it, well done!

Does it make sense that the answer was dragon fruit?

A17: Yes, it makes sense that the answer was dragon fruit, because it is a fruit that is purple on the outside and edible. It is not a berry, a grape, a plum, a fig, or a passion fruit, so it fits the criteria of the previous questions. It is also not a very common fruit, so it might be harder to guess than some other options.

Does GPT-4 conform to our expectations in conversational use?

- Shows signs of keeping track of conversational revealment
 - 20 questions, with detailed post mortem.
- More probing is needed: Does it start to forget context?
 - Do we need the “interaction DB” like in Macaw, that can bring up salient reminders that the model may have forgotten.
- Bing Chat is an example of conversational search using GPT-4
- How will the community develop, improve, measure, etc?
 - I’m confident we’ll figure it out.

Thanks!