snowflake®

How Agents and Humans Reason Over Document Collections

Łukasz Borchmann

snowflake®

서울 가서 김 서방 찾기

Łukasz Borchmann

‘Seoul’

서울 가서 김 서방 찾기

‘Kim’

서울 가서 김 서방 찾기

Kim 김
21.5%

Lee 이
14.7%

Park 박
8.4%

Choi 최
4.7%

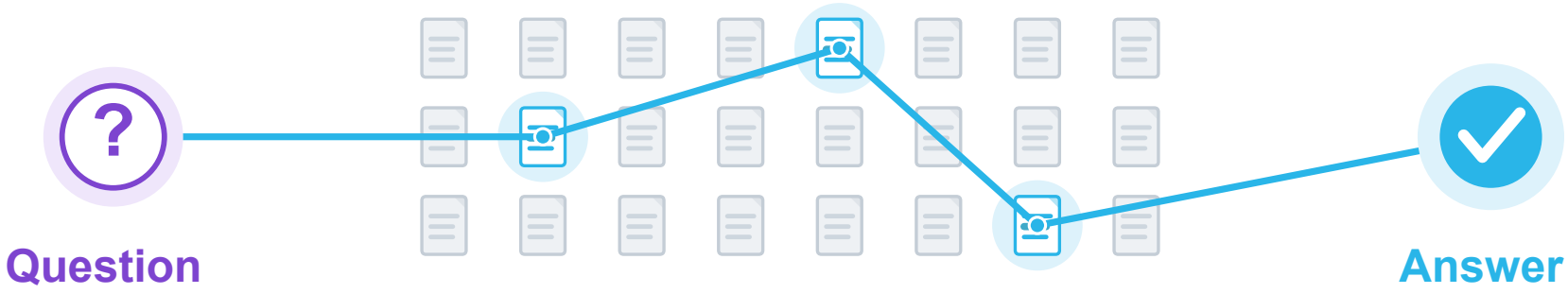
서울 가서 김 서방 찾기

‘Going to Seoul to look for Mr. Kim’

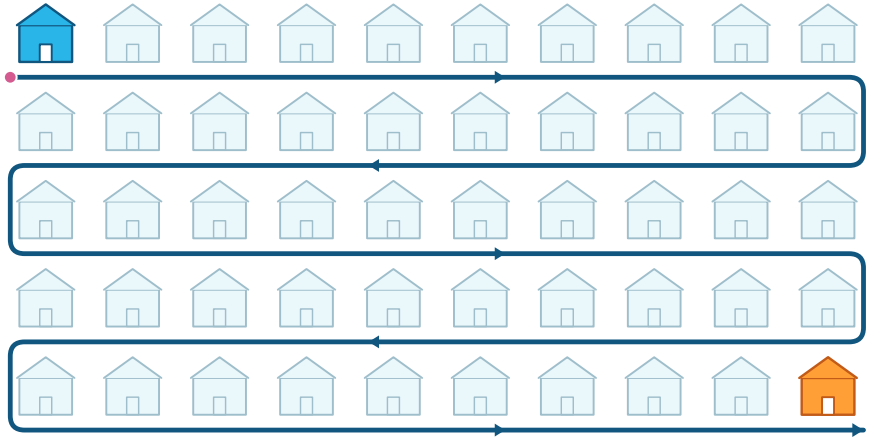
Question

Answer

Collection of documents



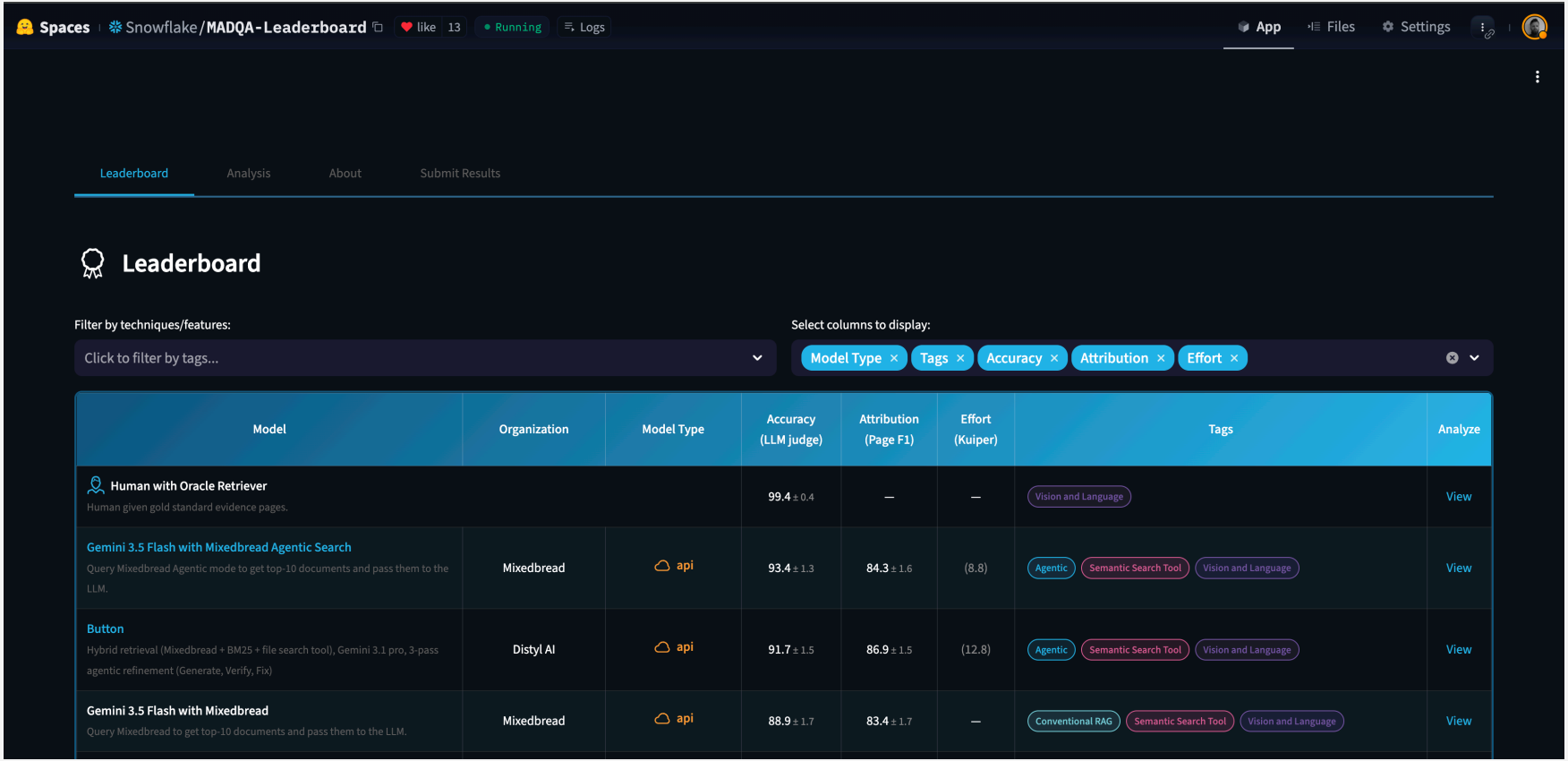
Collection of documents



Łukasz **Borchmann** • Jordy Van **Landeghem** • Michał **Turski**
Shreyansh **Padarha** • Ryan Othniel **Kearns** • Adam **Mahdi**
Niels **Rogge** • Clémentine **Fourrier** • Siwei **Han** • Huaxiu **Yao**
Artemis **Llabrés** • Yiming **Xu** • Dimosthenis **Karatzas**
Hao **Zhang** • Anupam **Datta**

MADQA Benchmark

https://huggingface.co/spaces/Snowflake/MADQA-Leaderboard

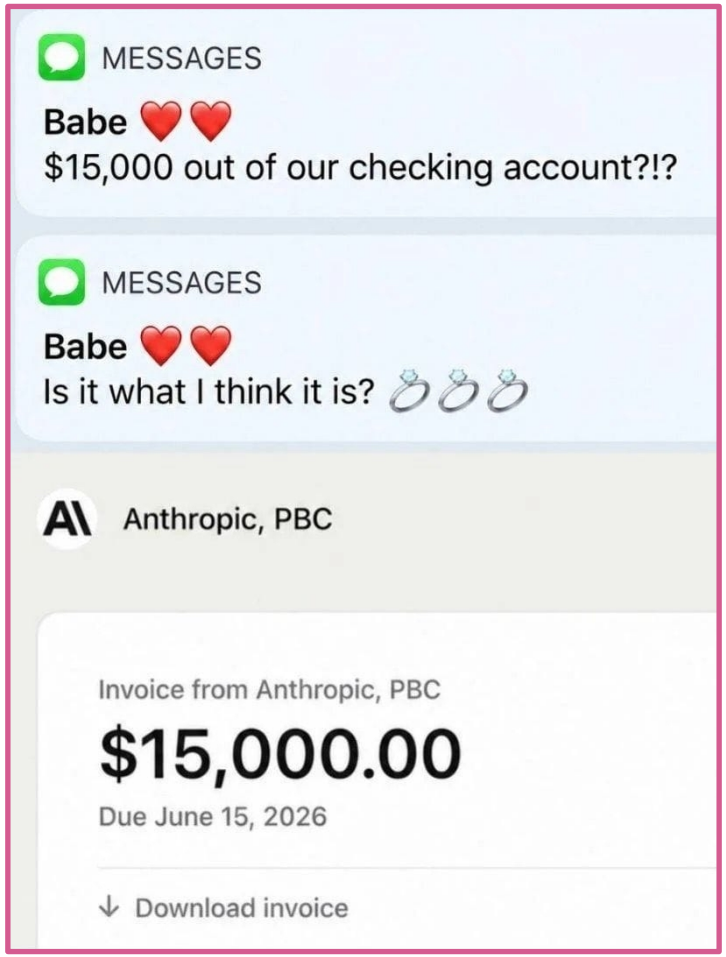


Why you should care

- How to build a good benchmark?
- How to compare agents which use variable compute?
- What are the specific failure modes of frontier models?

Test-time compute could buy higher accuracy. Can your agent be trusted to know when to spend it — and when to stop?

Nope.


MESSAGES
Babe ❤️❤️
$15,000 out of our checking account?!?
MESSAGES
Babe ❤️❤️
Is it what I think it is? 💍💍💍
Anthropic, PBC
Invoice from Anthropic, PBC
$15,000.00
Due June 15, 2026
Download invoice

The best LLM agents can now match human accuracy in document intelligence tasks.

The best LLM agents can now match human accuracy in document intelligence tasks.

But they solve problems 5 times less efficiently.

The best LLM agents can now match human accuracy in document intelligence tasks.

But they solve problems 5 times less efficiently.

And both humans and machines hit a performance ceiling.

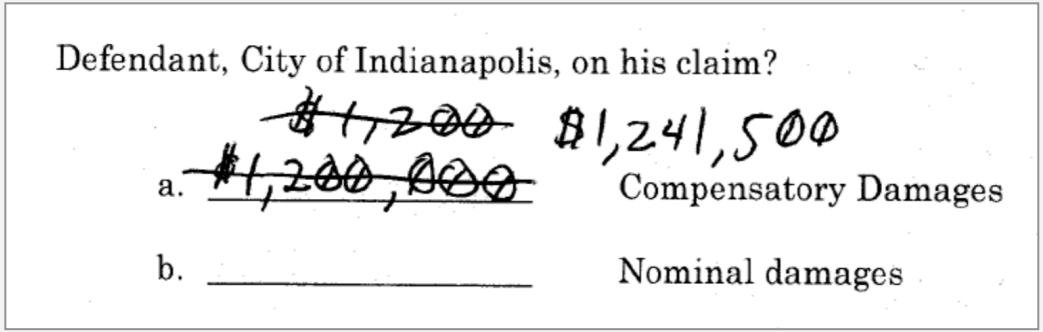
MADQA Benchmark

MADQA Benchmark. Collection of Documents

800 PDFs

Manually selected. Intentionally seeking **clusters of up to 30 related documents** (e.g., sequential reports or menus from different restaurants).

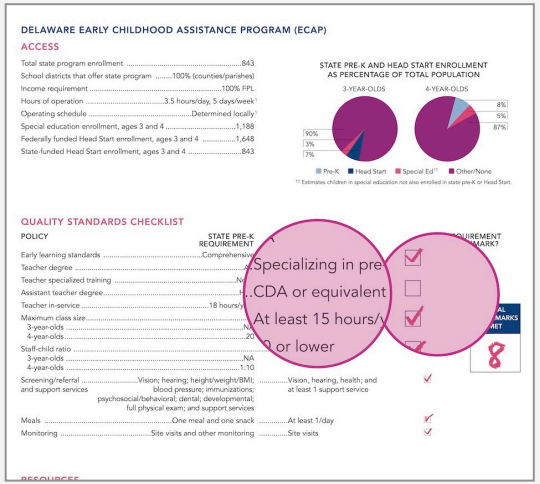
MADQA Benchmark. Collection of Documents



Defendant, City of Indianapolis, on his claim?

a. ~~$1,200~~ ~~$1,200,000~~ $1,241,500 Compensatory Damages

b. _____ Nominal damages



DELAWARE EARLY CHILDHOOD ASSISTANCE PROGRAM (ECAP)

ACCESS

Total state program enrollment 843
School districts that offer state program ..... 100% (counties/parishes)
Income requirement ..... 100% FPL
Hours of operation ..... 3.5 hours/day, 5 days/week[1]
Operating schedule ..... Determined locally[1]
Special education enrollment, ages 3 and 4 1,188
Federally funded Head Start enrollment, ages 3 and 4 1,648
State-funded Head Start enrollment, ages 3 and 4 843

STATE PRE-K AND HEAD START ENROLLMENT AS PERCENTAGE OF TOTAL POPULATION

3-YEAR-OLDS: 90%, 3%, 7%
4-YEAR-OLDS: 8%, 5%, 87%
Pre-K, Head Start, Special Ed[1], Other/None

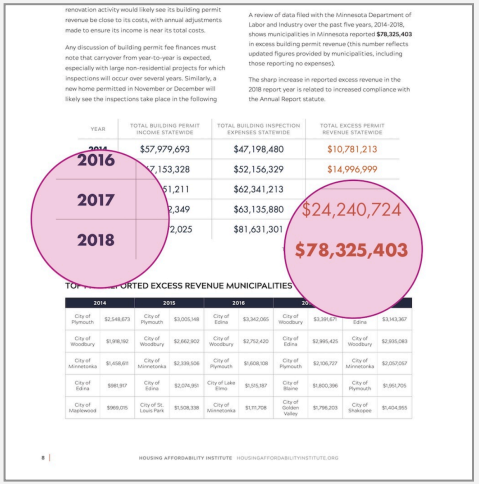
[1] Estimates children in special education not also enrolled in state pre-K or Head Start.

QUALITY STANDARDS CHECKLIST

POLICY	STATE PRE-K REQUIREMENT	MEETS BENCHMARKS?
Early learning standards	Comprehensive	
Teacher degree	A... Specializing in pre	✓
Teacher specialized training	N...	
Assistant teacher degree	B... CDA or equivalent	
Teacher in-service	18 hours/y...	✓
Maximum class size		
3-year-olds	N... At least 15 hours/y...	
4-year-olds	20 or lower	
Staff-child ratio		
3-year-olds	NA	
4-year-olds	1:10	
Screening/referral and support services	Vision; hearing; height/weight/BMI; blood pressure; immunizations; psychosocial/behavioral; dental; developmental; full physical exam; and support services	Vision, hearing, health; and at least 1 support service ✓
Meals	One meal and one snack	At least 1/day ✓
Monitoring	Site visits and other monitoring	Site visits ✓

8

RESOURCES



renovation activity would likely see its building permit revenue be close to its costs, with annual adjustments made to ensure its income is near its total costs.

Any discussion of building permit fee finances must note that carryover from year-to-year is expected, especially with large non-residential projects for which inspections will occur over several years. Similarly, a new home permitted in November or December will likely see the inspections take place in the following

A review of data filed with the Minnesota Department of Labor and Industry over the past five years, 2014-2018, shows municipalities in Minnesota reported **$78,325,403** in excess building permit revenue (this number reflects updated figures provided by municipalities, including those reporting no expenses).

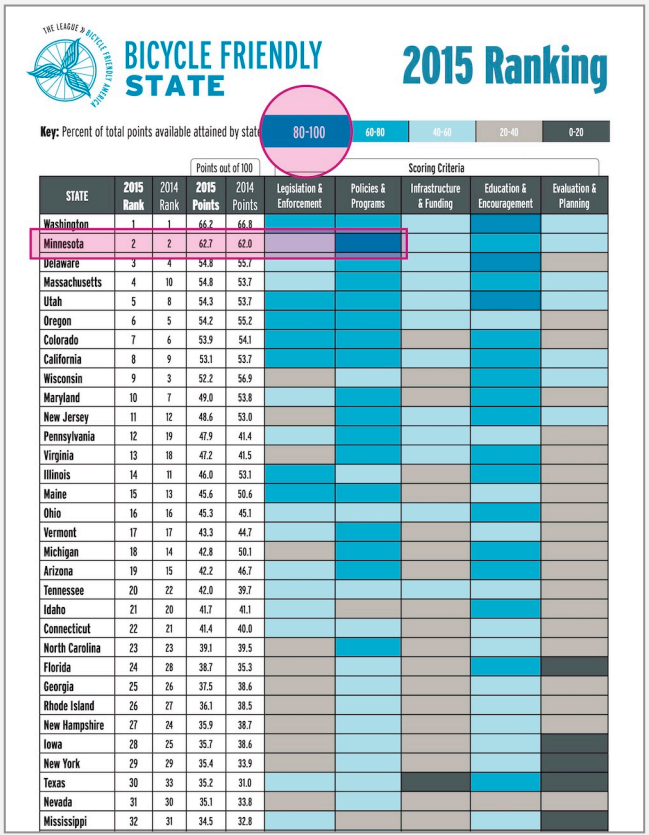
The sharp increase in reported excess revenue in the 2018 report year is related to increased compliance with the Annual Report statute.

YEAR	TOTAL BUILDING PERMIT INCOME STATEWIDE	TOTAL BUILDING INSPECTION EXPENSES STATEWIDE	TOTAL EXCESS PERMIT REVENUE STATEWIDE
2016	$57,979,693	$47,198,480	$10,781,213
2017	7,153,328	$52,156,329	$14,996,999
2018	1,211	$62,341,213	$24,240,724
	2,349	$63,135,880	
	2,025	$81,631,301	$78,325,403

TOP REPORTED EXCESS REVENUE MUNICIPALITIES

2014	2015	2016	2017	2018
City of Plymouth $2,548,673	City of Plymouth $3,005,148	City of Edina $3,342,065	City of Woodbury $3,281,012	City of Edina $3,143,367
City of Woodbury $1,918,192	City of Woodbury $2,662,902	City of Woodbury $2,752,420	City of Edina $2,995,425	City of Woodbury $2,935,083
City of Minnetonka $1,458,611	City of Minnetonka $2,339,506	City of Plymouth $1,608,108	City of Plymouth $2,106,727	City of Minnetonka $2,057,057
City of Edina $981,917	City of Edina $2,074,951	City of Lake Elmo $1,515,187	City of Blaine $1,800,396	City of Plymouth $1,951,705
City of Maplewood $969,015	City of St. Louis Park $1,508,338	City of Minnetonka $1,117,708	City of Golden Valley $1,796,203	City of Shakopee $1,404,955

8 | HOUSING AFFORDABILITY INSTITUTE HOUSINGAFFORDABILITYINSTITUTE.ORG



THE LEAGUE OF AMERICAN BICYCLIST'S BICYCLE FRIENDLY AMERICA

BICYCLE FRIENDLY STATE

2015 Ranking

Key: Percent of total points available attained by state: 80-100 | 60-80 | 40-60 | 20-40 | 0-20

			Points out of 100		Scoring Criteria				
STATE	2015 Rank	2014 Rank	2015 Points	2014 Points	Legislation & Enforcement	Policies & Programs	Infrastructure & Funding	Education & Encouragement	Evaluation & Planning
Washington	1	1	66.2	66.8					
Minnesota	2	2	62.7	62.0					
Delaware	3	4	54.8	55.7					
Massachusetts	4	10	54.8	53.7					
Utah	5	8	54.3	53.7					
Oregon	6	5	54.2	55.2					
Colorado	7	6	53.9	54.1					
California	8	9	53.1	53.7					
Wisconsin	9	3	52.2	56.9					
Maryland	10	7	49.0	53.8					
New Jersey	11	12	48.6	53.0					
Pennsylvania	12	19	47.9	41.4					
Virginia	13	18	47.2	41.5					
Illinois	14	11	46.0	53.1					
Maine	15	13	45.6	50.6					
Ohio	16	16	45.3	45.1					
Vermont	17	17	43.3	44.7					
Michigan	18	14	42.8	50.1					
Arizona	19	15	42.2	46.7					
Tennessee	20	22	42.0	39.7					
Idaho	21	20	41.7	41.1					
Connecticut	22	21	41.4	40.0					
North Carolina	23	23	39.1	39.5					
Florida	24	28	38.7	35.3					
Georgia	25	26	37.5	38.6					
Rhode Island	26	27	36.1	38.5					
New Hampshire	27	24	35.9	38.7					
Iowa	28	25	35.7	38.6					
New York	29	29	35.4	33.9					
Texas	30	33	35.2	31.0					
Nevada	31	30	35.1	33.8					
Mississippi	32	31	34.5	32.8					

MADQA Benchmark

QUESTION

What was the total excess permit revenue in MN for the 2014-2019 period?

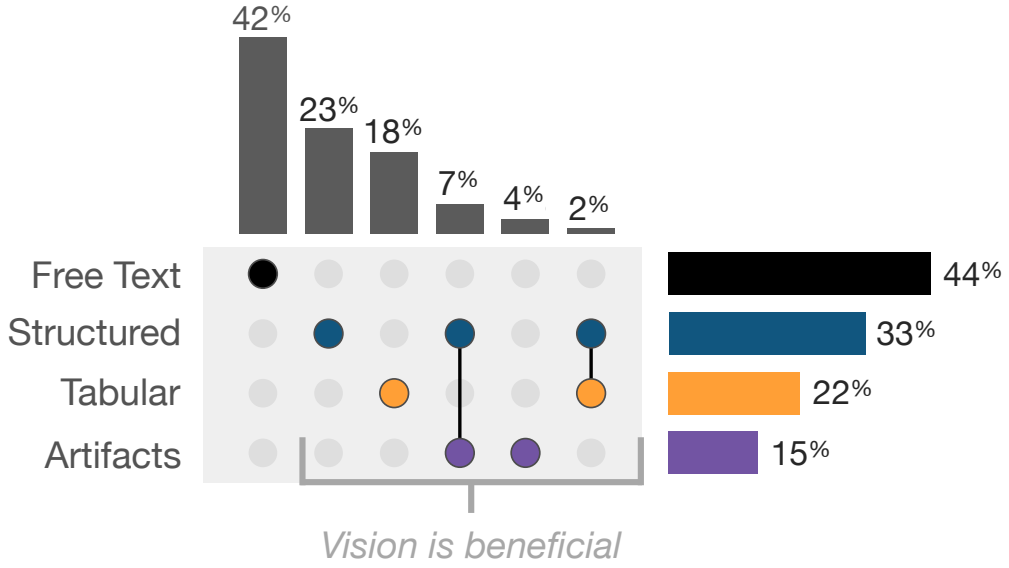
MADQA Benchmark. Questions over Document Collection

2,250 QA pairs

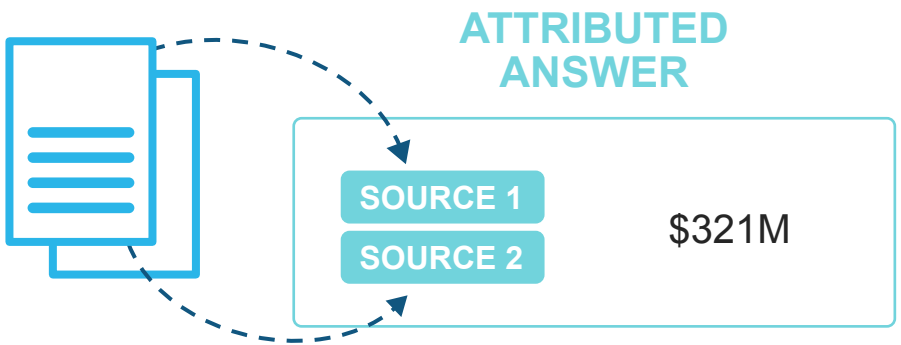
1,200 hours of professional annotation.

20% multi-hop.

More than 50% benefit from visual comprehension.



MADQA Benchmark. Answer with Source Attribution



MADQA Benchmark. Evaluation Angles

Correctness

Is the answer correct and actionable?

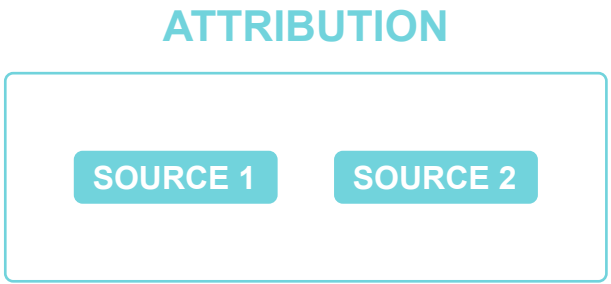
ANSWER

$321M

MADQA Benchmark. Evaluation Angles

Attribution

Is the answer well grounded?



MADQA Benchmark. Evaluation Angles

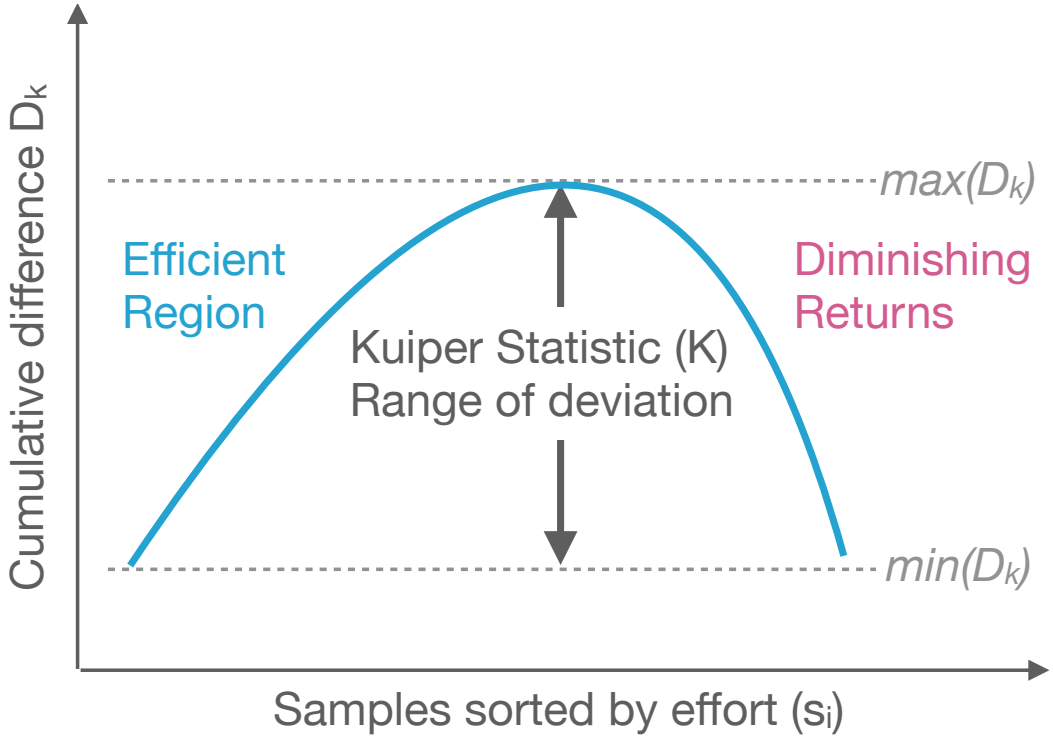
Efficiency

Is the effort agent invested justified?

MADQA Benchmark. Evaluation Angles

Efficiency

Is the effort agent invested justified?



MADQA Benchmark. Evaluation Angles

Efficiency

Is the effort agent invested justified?

TUTORIAL

Calibration and Bias in Algorithms, Data, and Models: a tutorial on metrics and plots for measuring calibration, bias, fairness, reliability, and robustness

Mark Tygert

2025 Tutorial

ICML 2025

MADQA Benchmark. Evaluation Angles

Correctness

Is the answer correct and actionable?

Attribution

Is the answer well grounded?

Efficiency

Is the effort the agent invested justified?

Findings

Simple Agentic Systems Can Outperform Strong, Static RAG, e.g.:

	Static RAG Service	Acc.	BM25 Agent Acc.	Δ
Gemini	*Gemini File Search*	78.6	82.2	**+3.6**
GPT	*OpenAI Assistants File Search*	50.0	67.8	**+17.8**

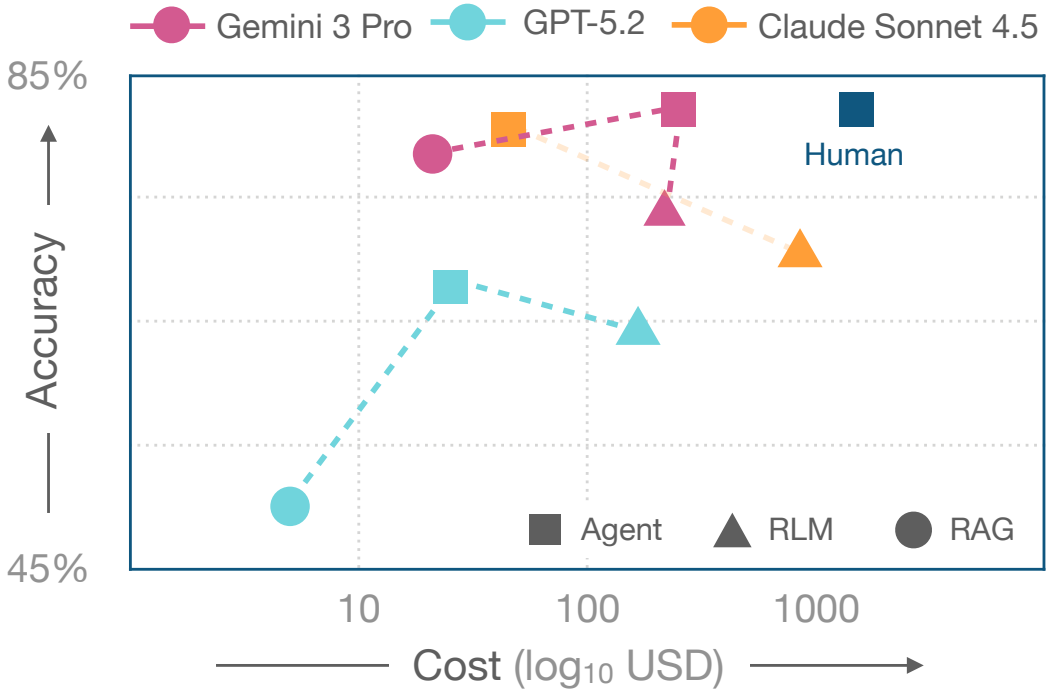
Model / Framework	Accuracy
Non-Agentic Systems	
Gemini 3 Pro File Search	**78.6** ± 2.2
Gemini 2.5 Flash File Search	71.8 ± 2.4
M3DocRAG	61.6 ± 2.6
GPT-5.2 (2024-08) HEAVEN	52.9 ± 2.7
GPT-5.2 (2025-12) File Search	50.0 ± 2.7
GPT-5 (2025-08) File Search	49.6 ± 2.7
GPT-4o (2024-08) HEAVEN	48.6 ± 2.7
GPT-5 Mini (2025-08) File Search	48.5 ± 2.7
ColBERTv2 + Llama-3.1-8B	40.2 ± 2.6
Agentic Systems	
Gemini 3 Pro BM25 Agent	**82.2** ± 2.0
Claude Sonnet 4.5 (2025-09) BM25 Agent	80.6 ± 2.1
GPT-5 (2025-08) BM25 Agent	77.7 ± 2.2
Gemini 3 Pro RLM	73.8 ± 2.3
Claude Agent Semtools	72.6 ± 2.4
Claude 4.5 Sonnet (2025-09) RLM	70.5 ± 2.4
Claude Haiku 4.5 (2025-10) BM25 Agent	68.2 ± 2.5
GPT-5.2 (2025-12) BM25 Agent	67.8 ± 2.5
GPT-5 Mini (2025-08) BM25 Agent	66.9 ± 2.5
GLM-4.6V BM25 Agent	66.1 ± 2.5
GPT-5.2 (2025-12) RLM	64.2 ± 2.6
MDocAgent	63.8 ± 2.6
Qwen3-VL (235B-A22B-Thinking) BM25 Agent	60.3 ± 2.6
GPT-4.1 (2025-04) BM25 Agent	60.0 ± 2.6
Gemini 2.5 Flash BM25 Agent	58.5 ± 2.6
GPT-5 Nano (2025-08) BM25 Agent	58.2 ± 2.6
Qwen3-VL (8B-Thinking) BM25 Agent	47.3 ± 2.7
GLM-4.6V Flash BM25 Agent	46.0 ± 2.7
GPT-4.1 Nano (2025-04) BM25 Agent	19.5 ± 2.1
Human Performance	
Human Oracle Retriever	**99.4** ± 0.4
Human BM25 Agent	82.2 ± 2.0

Findings

Retrieval Constraints are Essential for Cost-Effective Reasoning

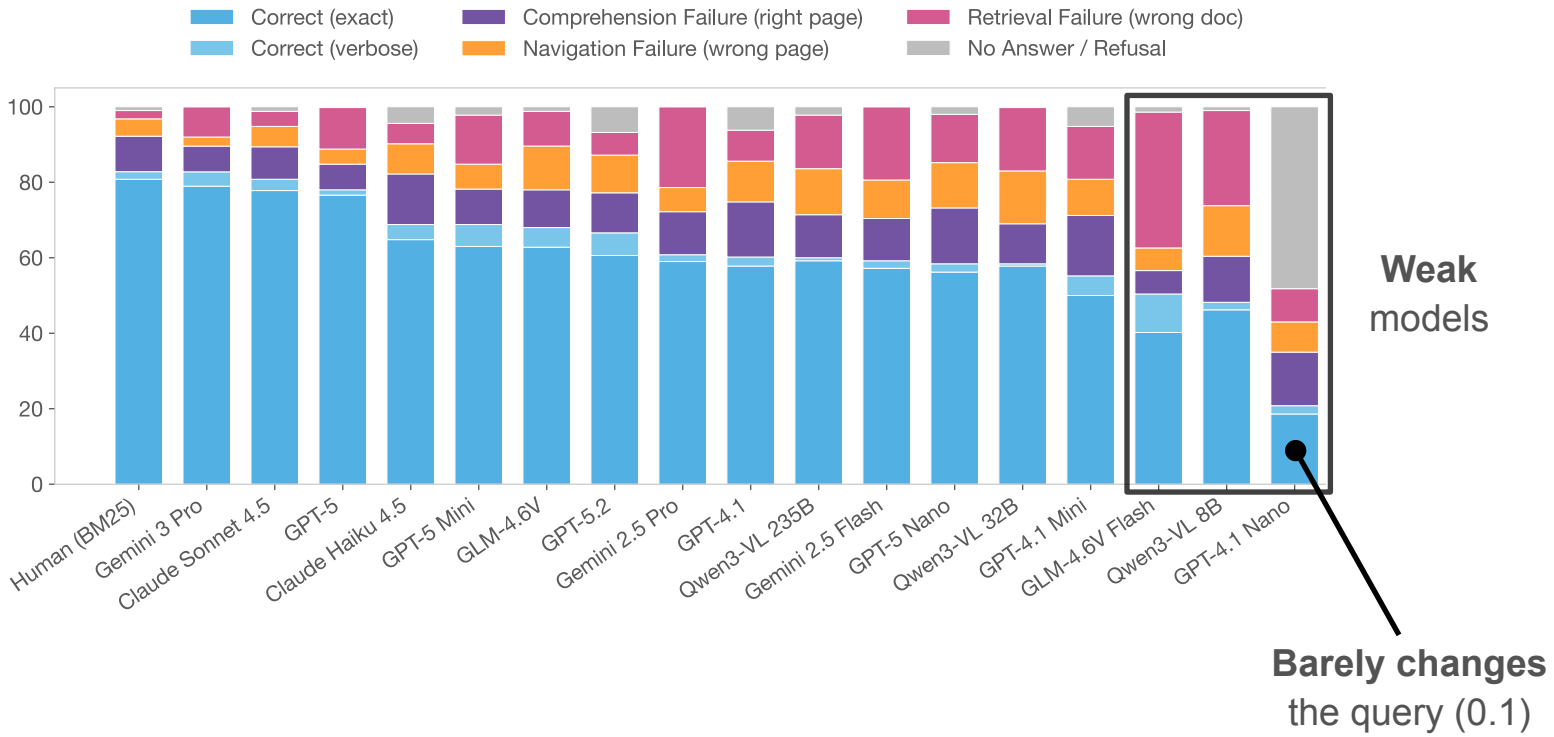
- The same **Claude** model used as RLM or BM25 Agent

	Cost	Acc.
RLM	$850	70%
BM25 Agent	$45	80%



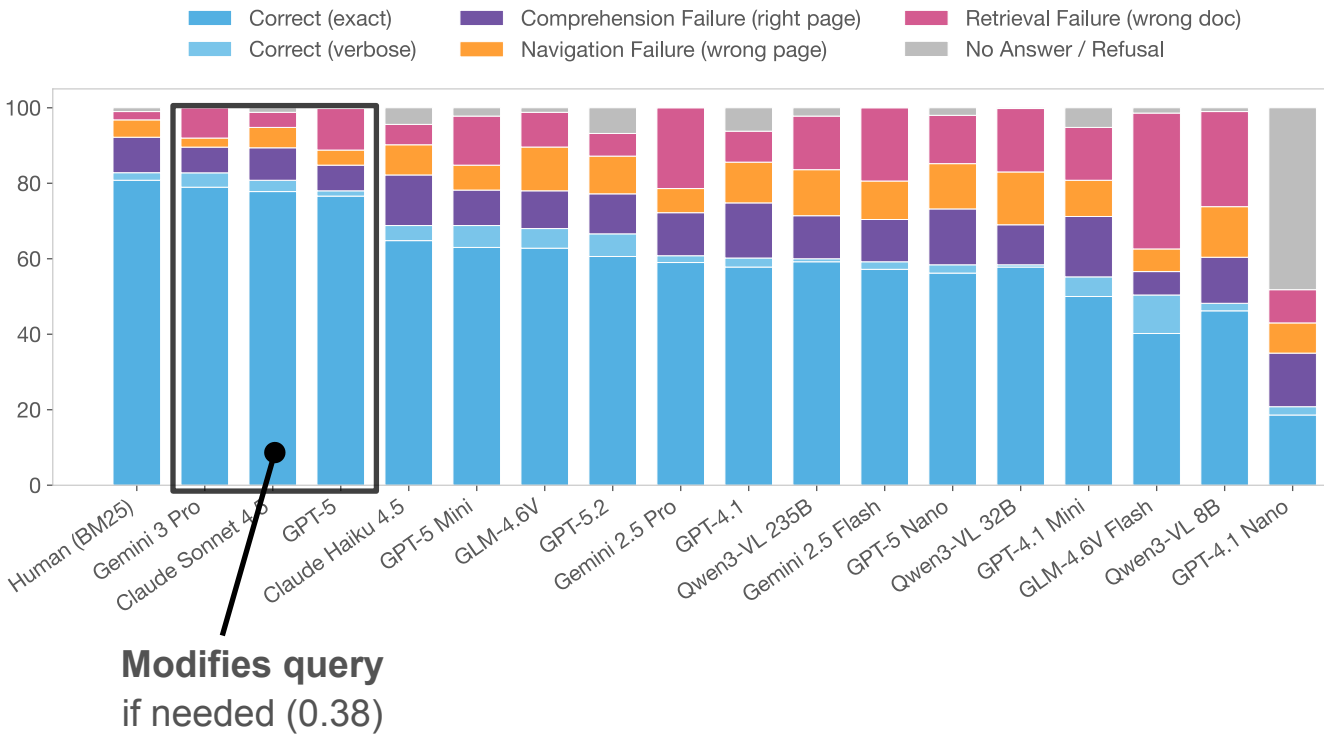
Findings

Failure Modes Differ Across Models. Query Reformulation Magnitude Predicts Success.



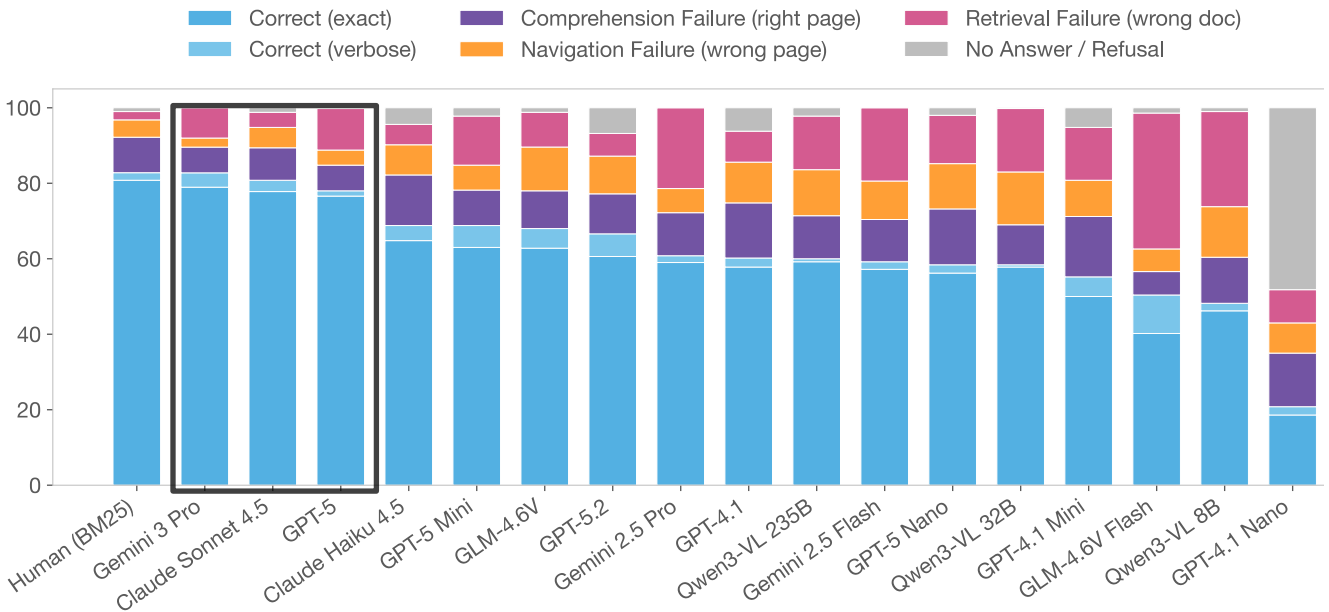
Findings

Failure Modes Differ Across Models. Query Reformulation Magnitude Predicts Success.



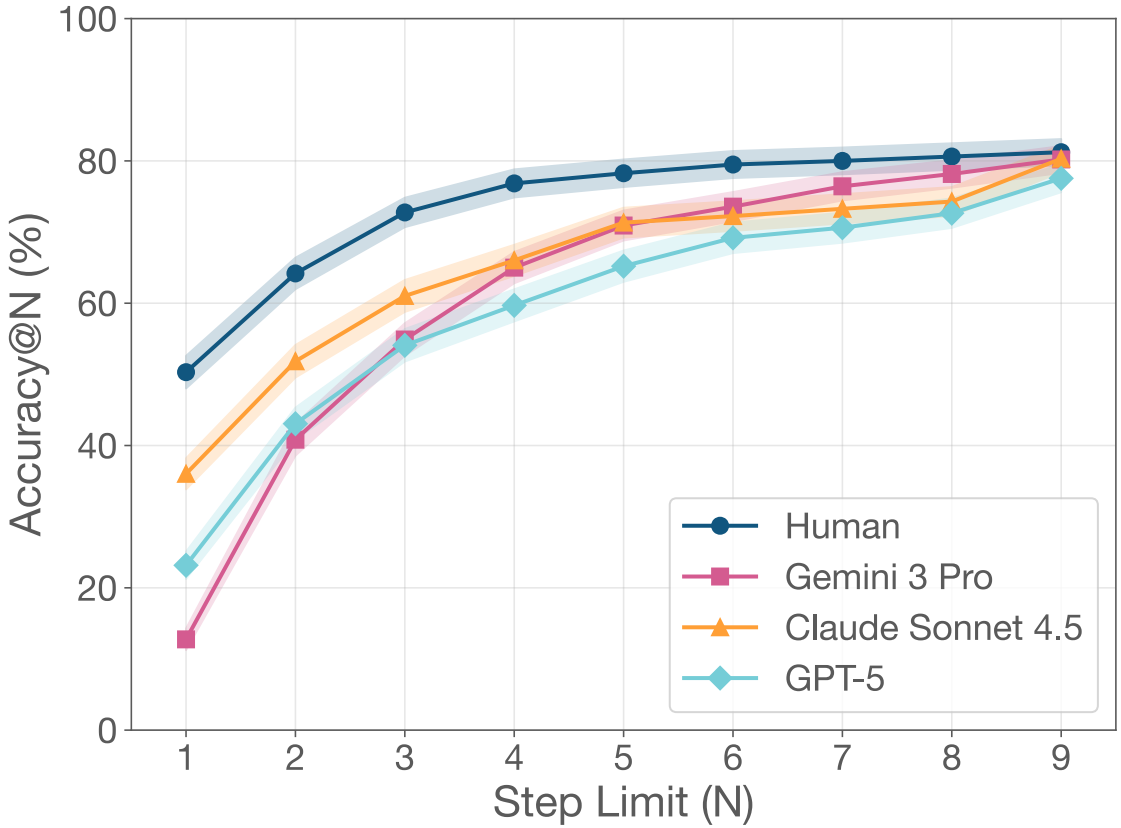
Findings

Failure Modes Differ Across Models. Query Reformulation Magnitude Predicts Success.



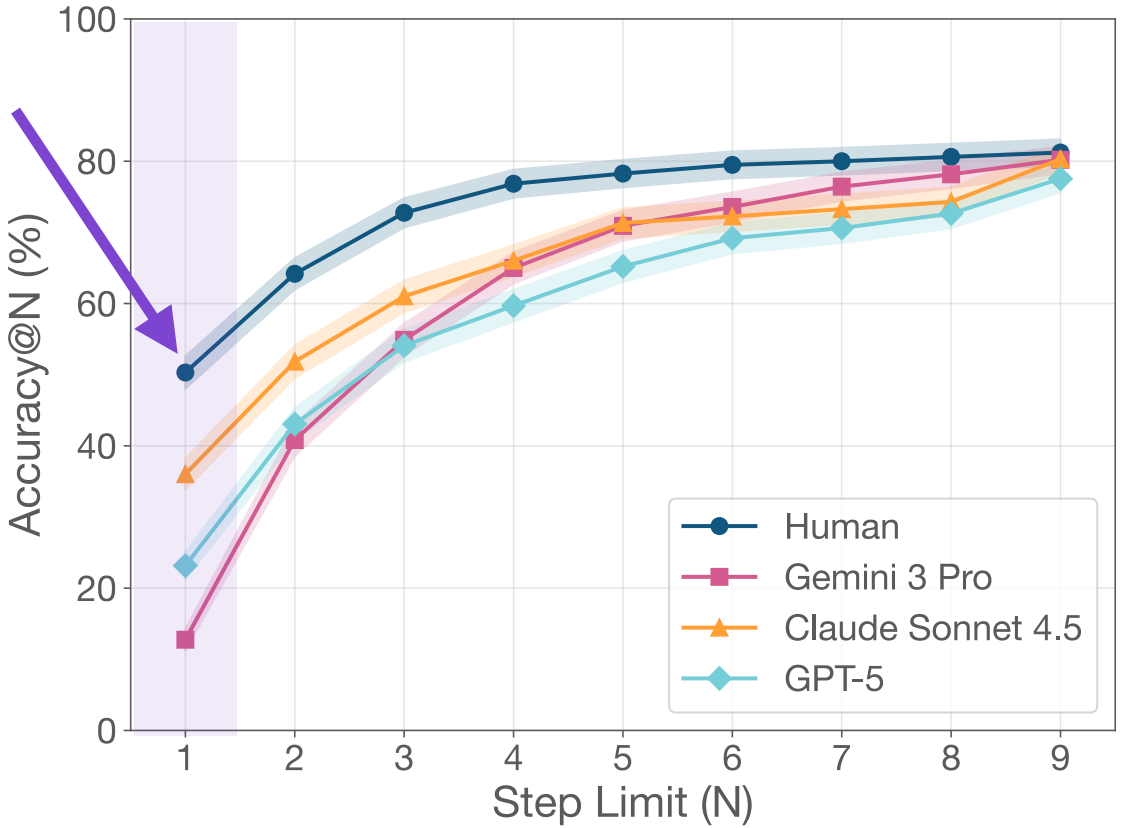
Findings

Illusion of Infinite Budget



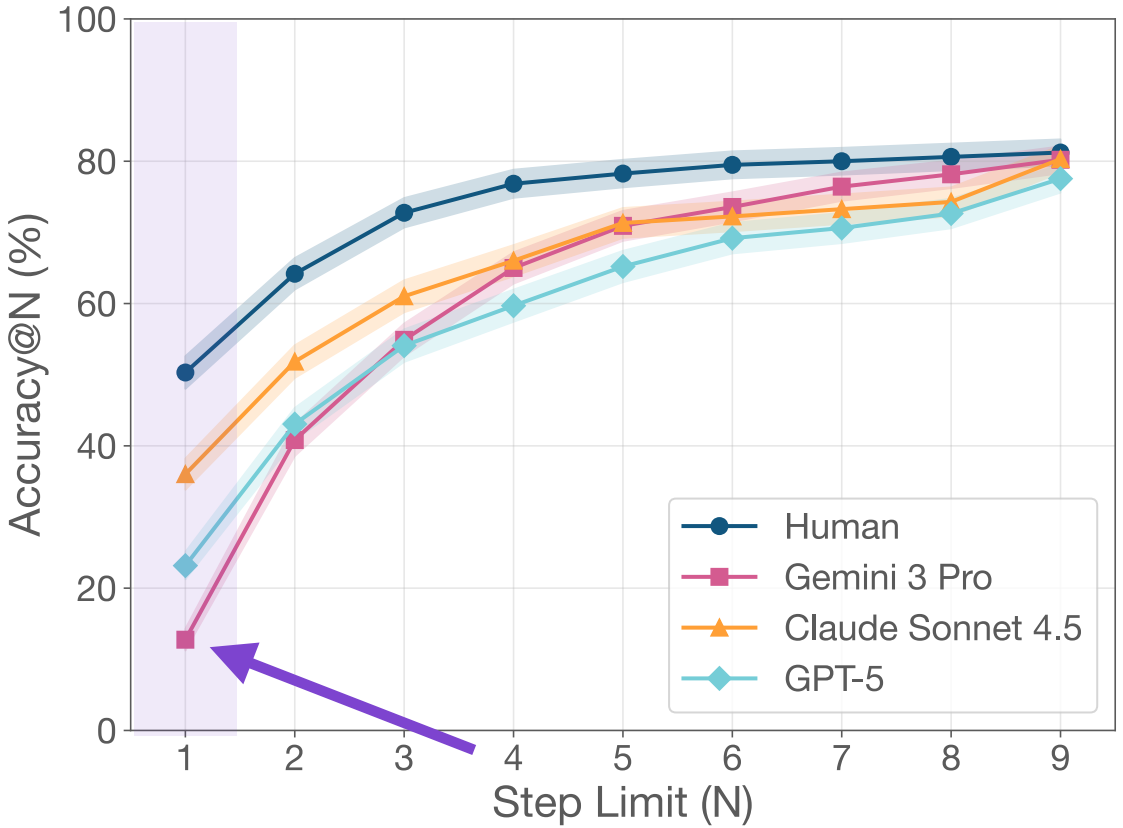
Findings

Illusion of Infinite Budget



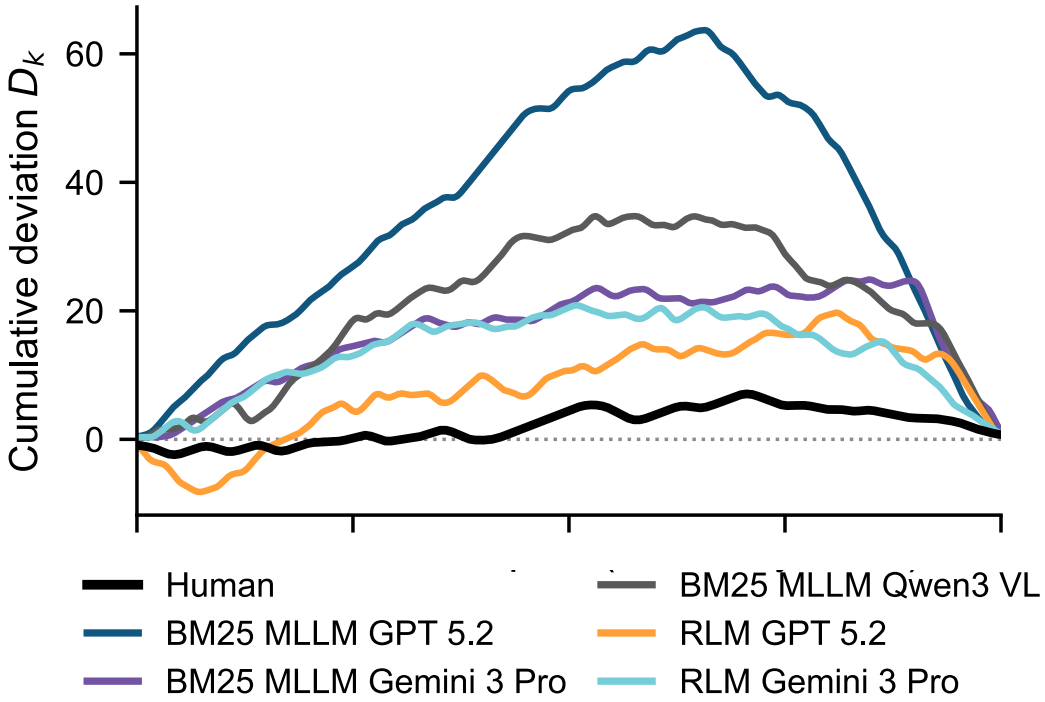
Findings

Illusion of Infinite Budget



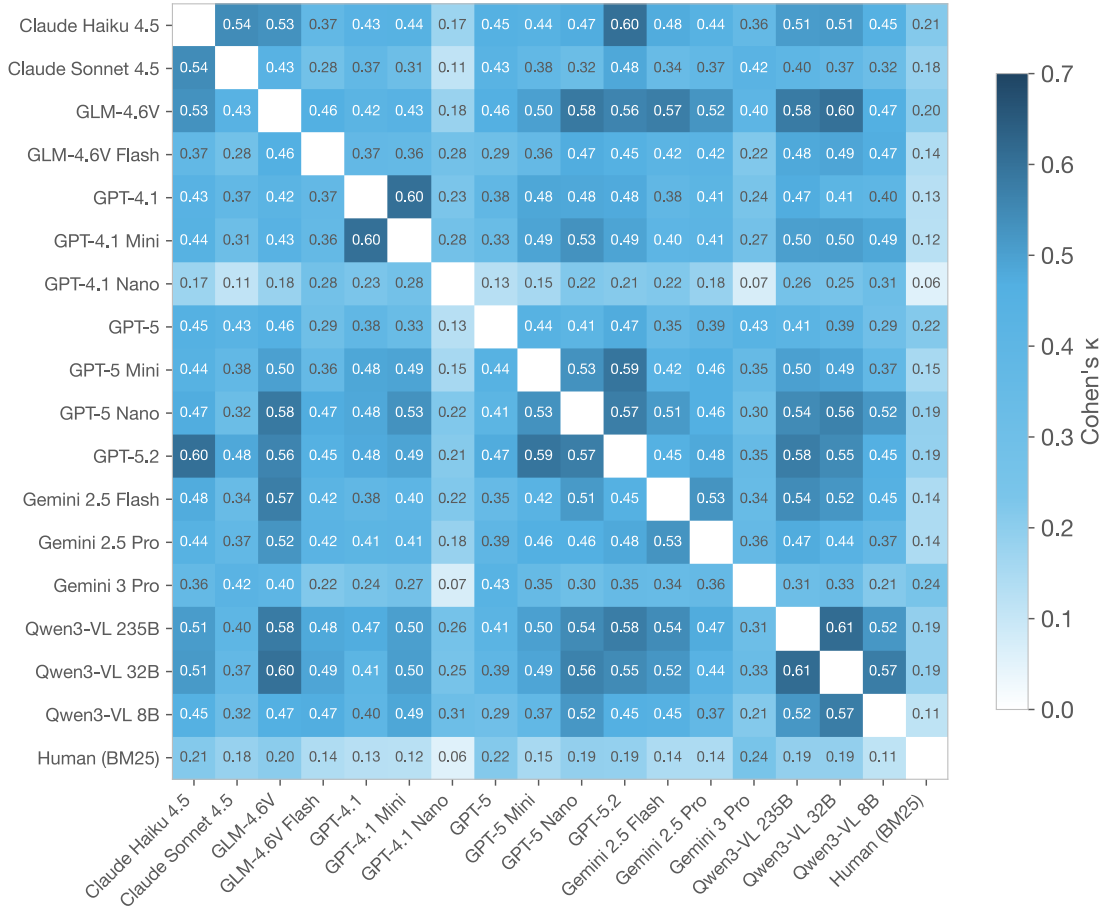
Findings

Humans
Calibrate
Effort
Better



Findings

Same Accuracy, Different Competencies



Findings

A human–agent pipeline could clear the ceiling neither reaches alone.

Why trust the readout and use our benchmark?

- We applied construct validity framework to validate the benchmark.

Why trust the readout and use our benchmark?

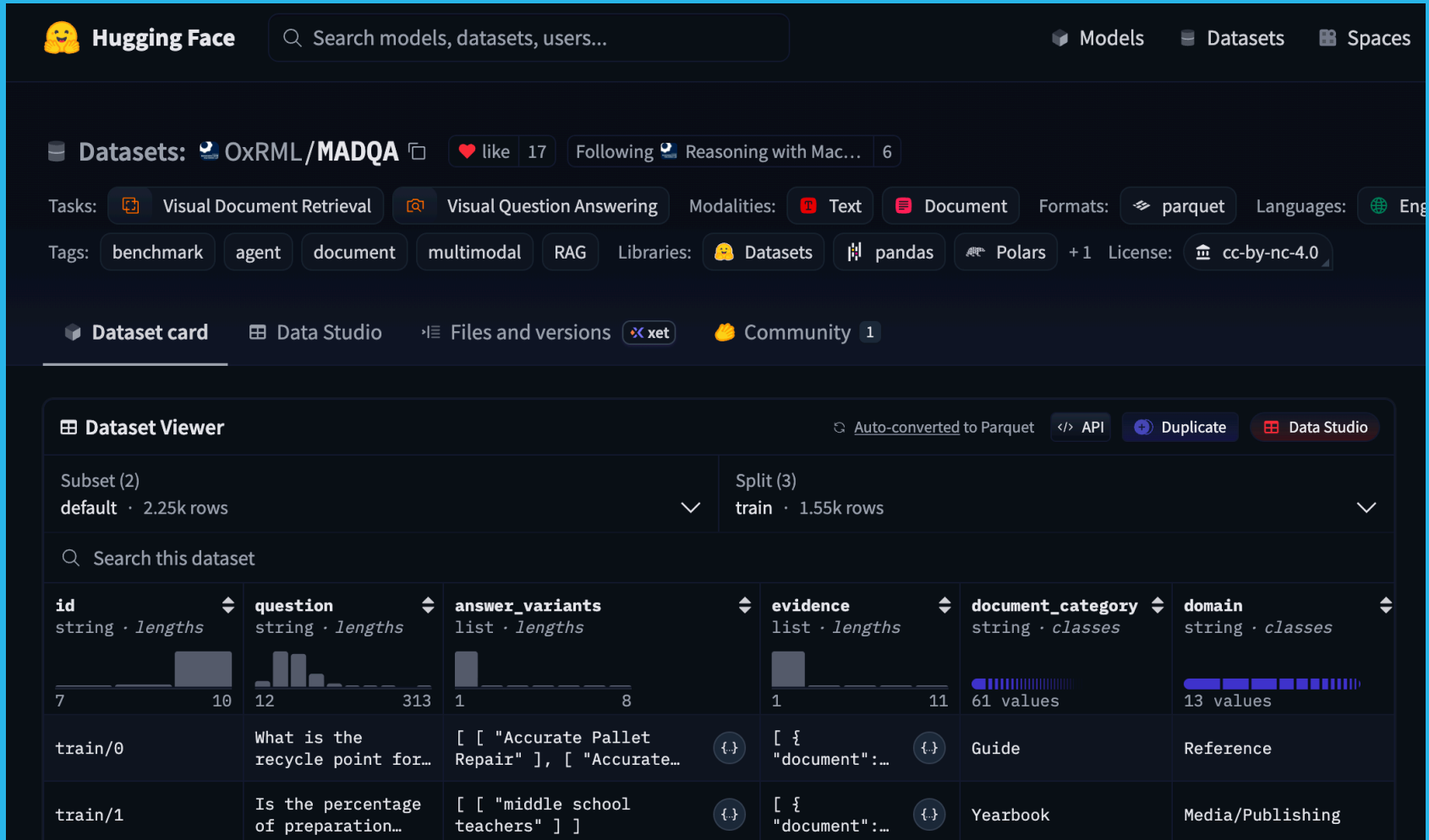
- We applied the construct validity framework to validate the benchmark. Checked, e.g.
  - Lexical Overlap vs. Reasoning

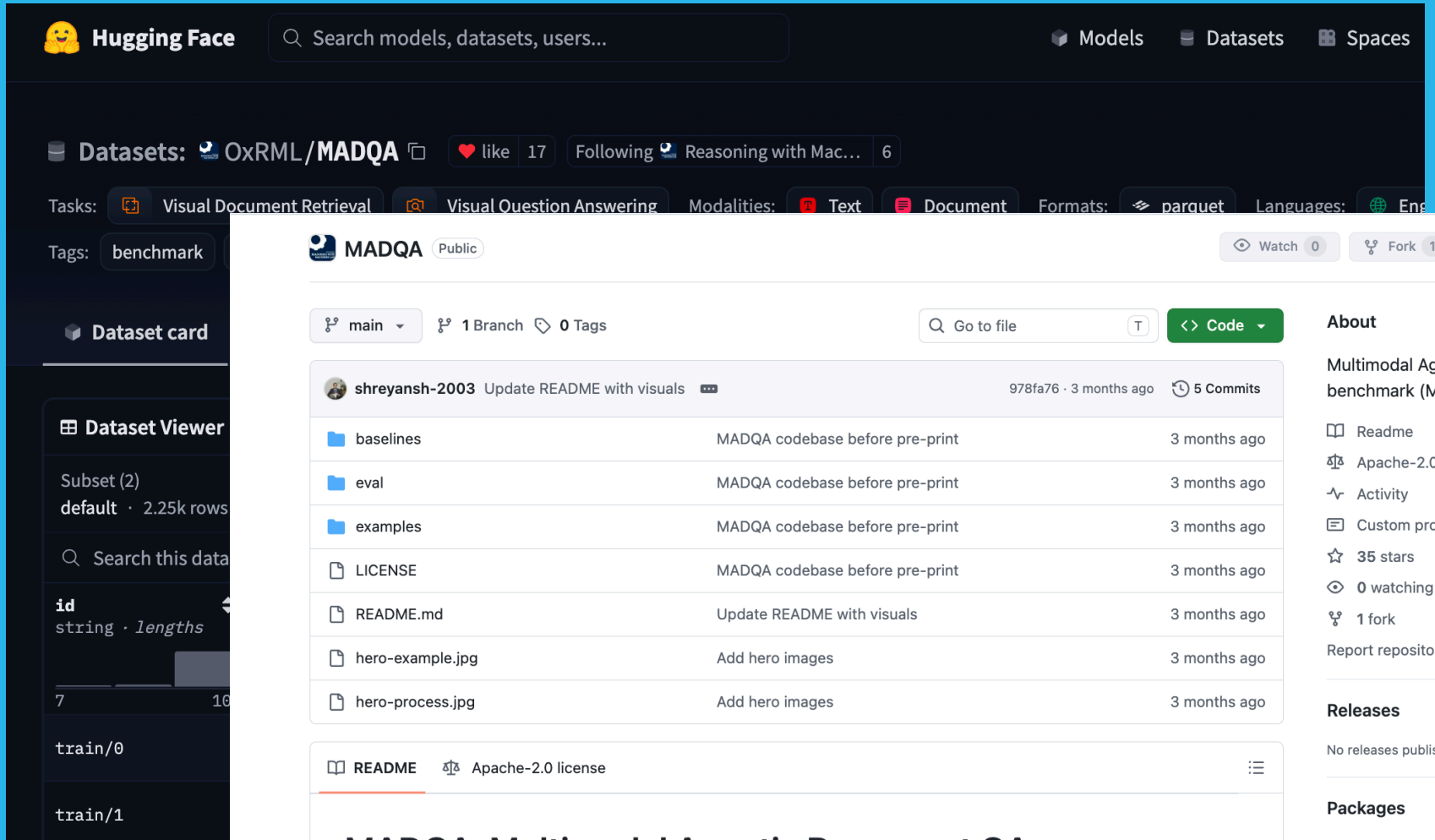
Why trust the readout and use our benchmark?

- We applied the construct validity framework to validate the benchmark. Checked, e.g.
  - Lexical Overlap vs. Reasoning
  - Parametric Knowledge vs. Grounding

Why trust the readout and use our benchmark?

- We applied the construct validity framework to validate the benchmark. Checked, e.g.
  - Lexical Overlap vs. Reasoning
  - Parametric Knowledge vs. Grounding
- It is not saturated. 80% achieved by the best models was our target.


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Datasets: OxRML/MADQA
like 17
Following Reasoning with Mac... 6
Tasks: Visual Document Retrieval
Visual Question Answering
Modalities: Text
Document
Formats: parquet
Languages: Eng
Tags: benchmark
agent
document
multimodal
RAG
Libraries: Datasets
pandas
Polars
+1
License: cc-by-nc-4.0
Dataset card
Data Studio
Files and versions
xet
Community 1
Dataset Viewer
Auto-converted to Parquet
API
Duplicate
Data Studio
Subset (2)
default · 2.25k rows
Split (3)
train · 1.55k rows
Search this dataset
id
string · lengths
7
10
question
string · lengths
12
313
answer_variants
list · lengths
1
8
evidence
list · lengths
1
11
document_category
string · classes
61 values
domain
string · classes
13 values
train/0
What is the recycle point for…
[["Accurate Pallet Repair"], ["Accurate…
[{ "document":…
Guide
Reference
train/1
Is the percentage of preparation…
[["middle school teachers"]]
[{ "document":…
Yearbook
Media/Publishing


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Datasets: OxRML/MADQA
like 17
Following Reasoning with Mac... 6
Tasks: Visual Document Retrieval
Visual Question Answering
Modalities: Text
Document
Formats: parquet
Languages:
Tags: benchmark
Dataset card
Dataset Viewer
Subset (2)
default · 2.25k rows
Search this data
id
string · lengths
7
10
train/0
train/1
MADQA Public
Watch 0
Fork 1
main
1 Branch
0 Tags
Go to file
T
Code
shreyansh-2003 Update README with visuals
978fa76 · 3 months ago
5 Commits
baselines MADQA codebase before pre-print 3 months ago
eval MADQA codebase before pre-print 3 months ago
examples MADQA codebase before pre-print 3 months ago
LICENSE MADQA codebase before pre-print 3 months ago
README.md Update README with visuals 3 months ago
hero-example.jpg Add hero images 3 months ago
hero-process.jpg Add hero images 3 months ago
README
Apache-2.0 license
About
Multimodal Agentic Docume
benchmark (MADQA)
Readme
Apache-2.0 license
Activity
Custom properties
35 stars
0 watching
1 fork
Report repository
Releases
No releases published
Packages

Leaderboard

Filter by techniques/features:

Click to filter by tags...

Select columns to display:

Model Type × Tags × Accuracy × Attribution × Effort ×

Model	Organization	Model Type	Accuracy (LLM judge)	Attribution (Page F1)	Effort (Kuiper)	Tags	Analyze
Human with Oracle Retriever Human given gold standard evidence pages.			99.4 ± 0.4	—	—	Vision and Language	View
Gemini 3.5 Flash with Mixedbread Agentic Search Query Mixedbread Agentic mode to get top-10 documents and pass them to the LLM.	Mixedbread	api	93.4 ± 1.3	84.3 ± 1.6	(8.8)	Agentic, Semantic Search Tool, Vision and Language	View
Button Hybrid retrieval (Mixedbread + BM25 + file search tool), Gemini 3.1 pro, 3-pass agentic refinement (Generate, Verify, Fix)	Distyl AI	api	91.7 ± 1.5	86.9 ± 1.5	(12.8)	Agentic, Semantic Search Tool, Vision and Language	View
Gemini 3.5 Flash with Mixedbread Query Mixedbread to get top-10 documents and pass them to the LLM.	Mixedbread	api	88.9 ± 1.7	83.4 ± 1.7	—	Conventional RAG, Semantic Search Tool, Vision and Language	View
Gemini 3 Pro with Mixedbread Query Mixedbread to get top-10 documents and pass them to the LLM.	Mixedbread	api	88.2 ± 1.7	82.2 ± 1.7	—	Conventional RAG, Semantic Search Tool, Vision and Language	View
Qwen3.6 35B A3B 8bit - Spectrum OCR + Agentic Retrieval Open-weight agentic system with vision	ARRAY Innovation	</> open-weight	87.7 ± 1.8	85.3 ± 1.6	16.7	Agentic, Vision and Language	View

LLM benchmarks typically value that Mr. Kim was found, and care less about the process.


LOTTE
LOTTE
MIDNIGHT
K-POP
한국 섬나당
한국석식당
한국식당
한국식당
서울 카페
명동거리
LOTTE
오뎅이
떡볶이 오뎅
김밥
지하철
김밥
원조 떡볶이
지하철
한국 식당
서울 카페
한국 식당
지하철
지하철
명동거리
Mr.
Kim?

THANK YOU

Strategic Navigation or Stochastic Search? How Agents and Humans Reason Over Document Collections

Poster
Today, 2:00 PM – 3:45 PM
HALL A #1903

Paper