UMassAmherst | Manning College of Information & Computer Sciences
Center for Intelligent Information Retrieval

Total Recall QA: A Verifiable Evaluation Suite for Deep Research Agents

Mahta Rafiee, Heydar Soudani, Zahra Abbasiantaeb, Mohammad Aliannejadi, Faegheh Hasibi, Hamed Zamani

Contributions

- Introduction of **Total Recall QA** (TRQA) task to address a proposed set of requirements and criteria for reliable evaluation of Deep Research Agents.
- An entity-centric framework for systematic generation of datasets for the TRQA task, leveraging a paired structured knowledge base and text corpus.
- Three datasets constructed using the proposed framework that require total recall retrieval and multi-step reasoning, including a synthetically generated e-commerce dataset to mitigate the effects of data contamination.
- Benchmark results and extensive analysis of intermediate retrieval and end-to-end performance of systems, showing commercial and open-source SOTA LLMs and deep research agents struggle to answer TRQA questions.

Challenges in Evaluating Deep Research Agents

We identify a taxonomy of **requirements** and optional **criteria** for reliable and reusable evaluation of DRAs:

- **Multi-source**: The answer to a deep research question must be a synthesis of information from multiple and sometimes many sources, involving complex reasoning and information retrieval processes.
- **Verifiability**: Metrics for evaluating the generated answers must be transparent, reliable, and verifiable.
- **Reproducibility**: Evaluation must be reproducible.
- **Retrieval Evaluation**: The evaluation methodology should provide a reliable, reproducible, and verifiable evaluation of intermediate retrieval steps.
- **Generalizability Evaluation**: The evaluation methodology should measure generalizability of DRAs across diverse aspects, including data contamination, memorization of popular information, and cross-domain generalization.

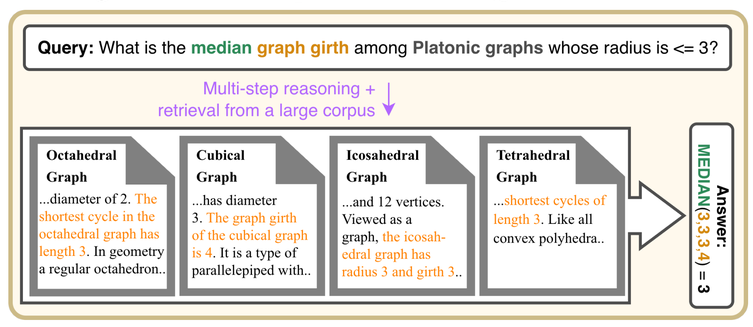
To address these requirements, inspired by TREC Total Recall Tracks, we introduce Total Recall QA:

Total Recall QA is a question answering task where accurate generation of the answer requires retrieving all relevant documents for a given question from a large corpus, as well as reasoning and synthesizing information across all relevant documents.

Table 1: Comparison of existing deep research benchmarks and TRQA across key requirements and criteria.

Benchmark	Multi-source	Verifiable	Reproducible	Retrieval	Generalizable
BrowseComp [39]	✓	✓	✗	✗	✗
BrowseComp-Plus [4]	✓	✓	✓	✓	✗
DeepResearchBench [8]	✓	✗	✗	✗	✓
DeepResearchGym [5]	✓	✗	✗	✗	✗
ReportBench [21]	✓	✗	✗	✗	✗
ScholarGym [32]	✓	✓	✓	✓	✗
DeepSearchQA [13]	✓	✓	✗	✓	✗
TRQA	✓	✓	✓	✓	✓

Figure 1: An example from the TRQA benchmark.



TRQA Data Generation Framework

Pipeline generates single-response queries with numerical answers for verifiability. Applied to 2 different sources to yield TRQA: Wikidata/Wikipedia + Synthetically generated e-commerce knowledge base and corpus

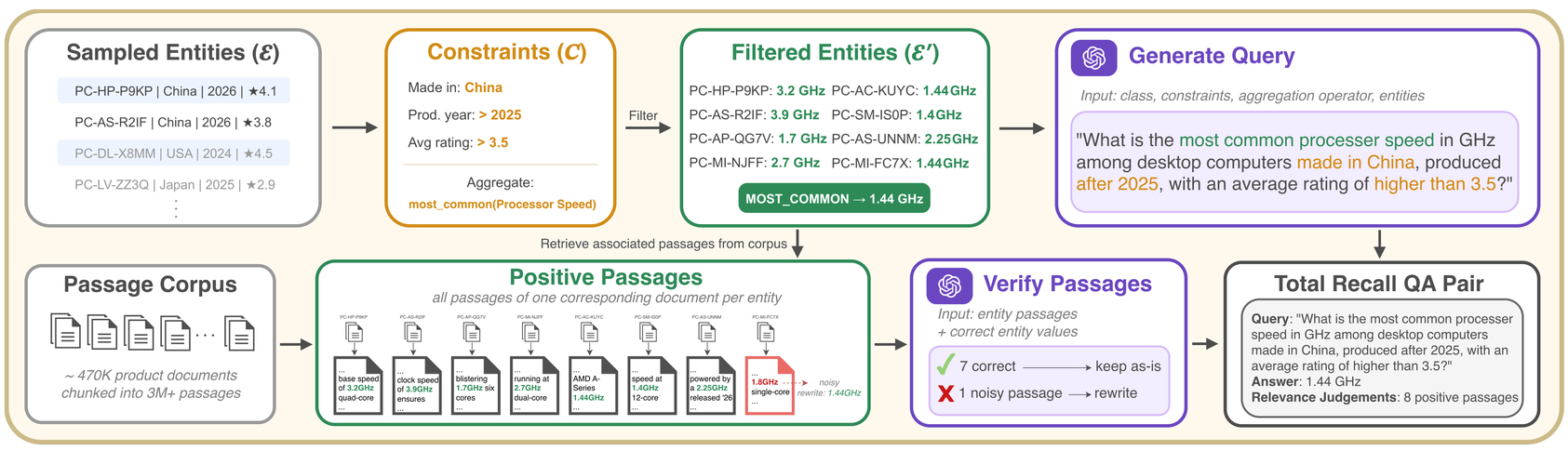


Table 2: Statistics of TRQA datasets. TRQA-Wiki1 and TRQA-Wiki2 share the same Wikipedia corpus.

Dataset	Test Queries	Validation Queries	Corpus Size (Full / Partial)	Relevant Entities per Query
TRQA-Wiki1	169	91	57.7M / 5.9M	12.78 ± 11.04
TRQA-Wiki2	1258	1083		9.01 ± 5.85
TRQA-Ecommerce	900	321	3.2M / -	22.22 ± 13.00

Table 3: Distribution of aggregation operations in TRQA.

Dataset	Count or Mode	Median or Mean	Min or Max	Sum	Time Ops	Distance	Percent
TRQA-Wiki1	47%	16%	16%	12%	5%	4%	—
TRQA-Wiki2	80%	6%	3%	2%	9%	—	—
TRQA-Ecommerce	52%	8%	8%	—	1%	—	31%

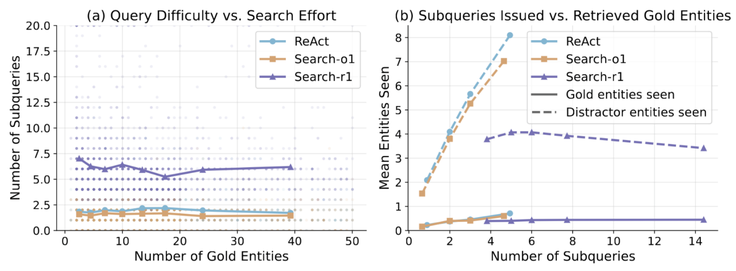
Benchmark Results and Findings

- **End-to-end evaluation metrics**: Exact Match over normalized answers + Soft Match (numeric answer correct if within X% of gold answer)
- **TRQA is challenging for DRAs**: consistent poor performance, underperforming GPT-5.2 on TRQA-Wiki1 and TRQA-Wiki2.
- **Impact of data contamination**: LLM-only results drop significantly on TRQA-Ecommerce.

Table 5: End-to-end performance of models. For DRAs, superscript $^{\Delta}$ indicates a statistically significant difference according to the McNemar test ($p < 0.05$) compared to E5 & Qwen2.5-7B. GPT-5.2 significantly outperforms other LLMs for all metrics.

	TRQA-Wiki1					TRQA-Wiki2					TRQA-Ecommerce				
	EM	$SM_{10\%}$	$SM_{20\%}$	$SM_{50\%}$	$SM_{90\%}$	EM	$SM_{10\%}$	$SM_{20\%}$	$SM_{50\%}$	$SM_{90\%}$	EM	$SM_{10\%}$	$SM_{20\%}$	$SM_{50\%}$	$SM_{90\%}$
LLMs															
Qwen2.5-7b-inst	2.36	4.73	7.10	14.20	21.89	2.46	10.65	11.69	19.48	29.57	1.00	2.67	4.56	11.33	26.89
Claude Sonnet 4.5	23.07	46.74	54.43	69.82	75.14	7.79	16.45	18.83	25.75	35.37	0.67	1.44	1.67	2.67	4.44
DeepSeek V3.2	24.34	42.11	48.68	61.84	68.42	7.23	16.38	19.87	28.30	38.79	1.88	3.77	6.44	12.55	28.00
Qwen3 235B A22B	16.88	29.38	36.25	56.25	66.88	8.03	18.84	21.07	33.23	48.65	2.56	5.22	**8.67**	**17.00**	34.89
GPT-4o	13.61	27.22	34.91	54.44	65.66	6.92	17.65	19.79	29.89	47.54	2.44	4.22	7.00	15.33	26.11
GPT-5.2	**30.18**	**53.85**	**64.50**	**75.74**	**82.84**	**10.33**	**21.70**	**25.91**	**40.70**	60.49	1.67	3.44	5.56	12.22	22.67
Deep Research Agents, Qwen2.5 7B, E5															
ReAct	5.33	9.47	15.38	26.63 $^{\Delta}$	37.28 $^{\Delta}$	5.32 $^{\Delta}$	15.42 $^{\Delta}$	16.37 $^{\Delta}$	23.84 $^{\Delta}$	38.55 $^{\Delta}$	0.88	2.44	3.88	8.00 $^{\Delta}$	17.22 $^{\Delta}$
Search-o1	11.24 $^{\Delta}$	18.34 $^{\Delta}$	24.85 $^{\Delta}$	39.05 $^{\Delta}$	56.80 $^{\Delta}$	6.12 $^{\Delta}$	17.48 $^{\Delta}$	19.95 $^{\Delta}$	29.09 $^{\Delta}$	50.31 $^{\Delta}$	1.55	3.77	6.22 $^{\Delta}$	13.22 $^{\Delta}$	29.22 $^{\Delta}$
Search-R1	10.65 $^{\Delta}$	20.12 $^{\Delta}$	23.67 $^{\Delta}$	43.20 $^{\Delta}$	58.58 $^{\Delta}$	4.85 $^{\Delta}$	14.47 $^{\Delta}$	16.53 $^{\Delta}$	26.23 $^{\Delta}$	55.25 $^{\Delta}$	2.55	4.77 $^{\Delta}$	7.22 $^{\Delta}$	14.55 $^{\Delta}$	33.44 $^{\Delta}$
Single Step Retriever (K=50)															
E5 & Qwen2.5-7b	3.55	9.46	11.83	17.75	28.40	3.42	7.00	8.82	14.47	24.32	1.56	2.33	4.78	9.22	21.33
E5 & GPT-5.2	**30.18**	48.52	52.66	66.27	74.56	9.46	20.27	24.32	38.63	**67.81**	**2.89**	**5.33**	7.56	16.56	**37.89**
Oracle & Claude 4.5	50.30	63.91	68.05	79.88	85.21	25.28	33.47	39.75	63.83	84.50	4.11	9.44	14.56	31.44	64.56
Oracle & GPT-5.2	56.80	72.19	76.92	84.02	89.35	27.11	36.33	43.48	68.36	92.61	5.56	9.67	14.78	29.33	67.22

- Average number of sub-queries issued stays the same across models regardless of the number of original query's gold entities.
- DRAs cannot effectively issue queries for finding relevant docs: as agents issue more sub-queries, they retrieve more distracting entities, few new gold entities.



(a): Avg # of sub-queries issued plotted against # of query gold entities

(b): Avg # of newly retrieved gold and distracting entities as more sub-queries are issued.

All models have extremely poor recall, but even with oracle retrieval, SOTA LLMs struggle. Oracle error analysis shows over 90% errors are caused by reasoning failure.

- Parametric Knowledge Bias: model generates same answer with and without oracle docs.
- Reasoning Error: model generates different incorrect answer.

DRA Model	TRQA-Wiki1	TRQA-Wiki2	TRQA-Ecommerce
ReAct	2.69	2.63	13.10
Search-o1	2.49	2.28 $^{\nabla}$	7.07 $^{\nabla}$
Search-R1	**3.27**	**3.37**	**13.95**

Table 6: Recall (%) of all intermediate retrieval steps (full corpus) for DRAs.

Model	Parametric Knowledge Bias	Reasoning Errors
Oracle & Claude 4.5	4.1%	95.9%
Oracle & GPT-5.2	9.1%	90.9%

Table 7: Distribution of error types by models with Oracle retrieval on all TRQA datasets combined.