UC San Diego

SourceBench: Can AI Answers Reference Quality Web Sources?

NeurIPS 2026 | Evaluations & Datasets Track

Hexi Jin · Stephen Liu · Yuheng Li · Simran Malik · Yiying Zhang University of California San Diego

66%

of exactly-correct AI answers cited at least one source with a severe failure on a source-quality dimension

“Severe failure” = a score of 2 or below on the 1–5 scale, on any of six dimensions: neutrality, freshness, ownership, author accountability, domain authority, layout clarity.

What existing benchmarks measure

Benchmark	What it tests	How
BrowseComp	Finding hard-to-find facts	Multi-step search questions; checks the short final answer
BrowseComp-Plus	Search over a fixed corpus	Compares retrievers and agents on evidence recall and accuracy
ScholarQABench, ResearchQA	Long-form research synthesis	Rubrics for coverage and correctness, plus citation checks
DeepResearch Bench	Report quality and citation support	RACE scores the report; FACT checks that cited pages support the claims

Was the information found? Is the report good? Do the citations support the claims?

Citation support is not source quality

Citation support (measured today)

Does this page support this sentence?

Passes when the page states the claim.

Source quality (SourceBench)

Is this page worth trusting and checking?

Accurate, neutral, current, accountable, readable.

A page can state the claim and still be outdated, promotional or anonymous. Content optimised for generative engines (GEO) is written to be cited.

What SourceBench measures

Dimension	Metric	What it measures
Content Quality	CR · Content Relevance	Does the page answer the user's question?
	FA · Factual Accuracy	Are its claims accurate and verifiable?
	NE · Neutrality	Does it present information objectively?
Source Credibility	FR · Freshness	Is the information current and maintained?
	OA · Ownership	Is the organization behind the site clear?
	AA · Author Accountability	Is the author identifiable and accountable?
	DA · Domain Authority	Is the source domain reputable?
Layout	LC · Layout Clarity	Is the page easy to read and verify?

From query to score

100 queries
Five intents, 20 each: shopping, multi-hop, factual, argumentative, social

→

12 systems
8 LLMs with search, Google top-5, 3 AI search tools

→

Cited URLs
Page text and metadata fetched for every source shown to the user

→

3 judge calls
Content, metadata, layout: each with its rubric and human-scored examples

→

8 scores
3,996 scored source instances, each with a short rationale

Only the sources a system finally shows the user are scored, not every page it visited along the way.

A judge calibrated against people

141

human-annotated pages

6 supply the frozen few-shot examples; the other 135 are a held-out validation set

0.75

mean absolute error

Qwen-Plus against human scores on the 1–5 scale, over the 135 held-out pages

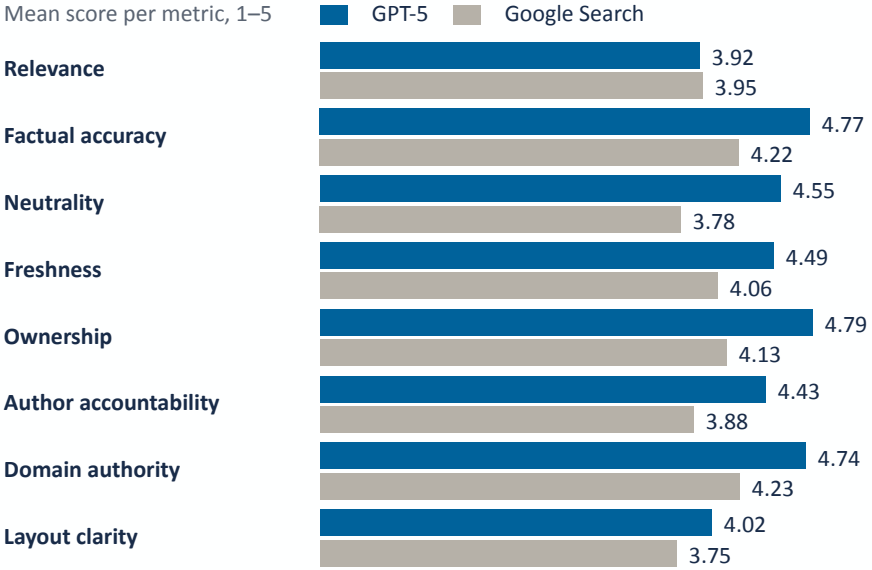
0.74

with an open-weights judge

Qwen3-235B-A22B-Instruct-2507 under identical prompts

Qwen-Plus is not one of the systems under test, and the pipeline runs end to end without a hosted API.

GPT-5 leads — but not on relevance



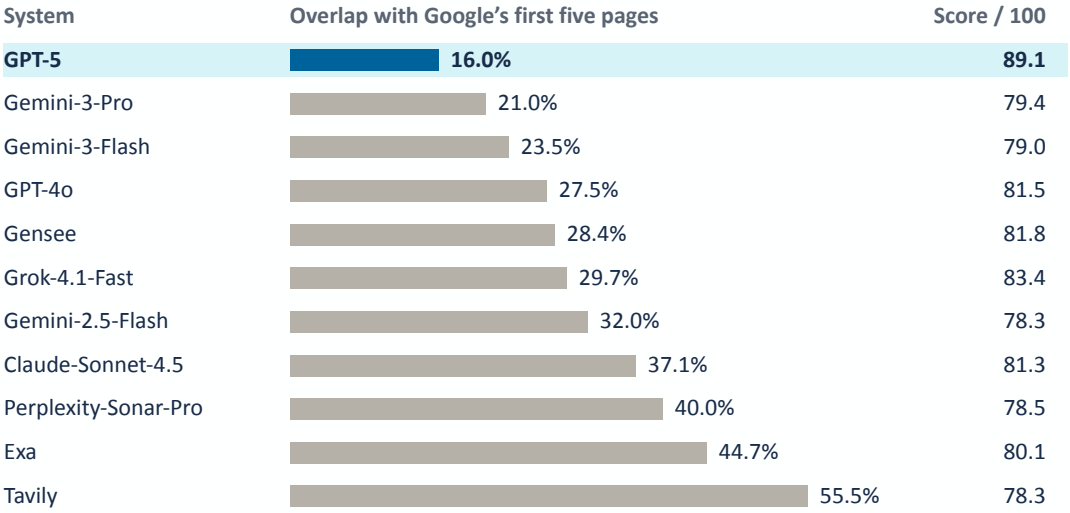
GPT-5 finishes first overall (89.1), ahead of Grok-4.1-Fast (83.4).

But Google edges it on relevance: 3.95 versus 3.92.

GPT-5 wins the other seven. The widest gaps are neutrality (+0.77) and ownership (+0.66): trust, not relevance.

16.0%

of GPT-5's cited sources appear in the first five pages of Google results — the lowest overlap of any system tested



System	Overlap with Google's first five pages	Score / 100
GPT-5	**16.0%**	**89.1**
Gemini-3-Pro	21.0%	79.4
Gemini-3-Flash	23.5%	79.0
GPT-4o	27.5%	81.5
Gensee	28.4%	81.8
Grok-4.1-Fast	29.7%	83.4
Gemini-2.5-Flash	32.0%	78.3
Claude-Sonnet-4.5	37.1%	81.3
Perplexity-Sonar-Pro	40.0%	78.5
Exa	44.7%	80.1
Tavily	55.5%	78.3

Retrieval substitutes for reasoning

Weighted source quality, out of 100

	Weaker retriever (Tavily)	Stronger retriever (Gensee)
Non-reasoning model	79.0	85.2
Reasoning model	85.6	83.3

One base model (DeepSeek-V3.2), two retrievers, 19 queries; essentially the same source quality either way!

Switching reasoning on helps for 14 of 19 queries with the weaker retriever, but only 5 of 19 with the stronger one. For source quality, reasoning is partly compensating for retrieval.

Limitations

The judge has error

MAE 0.75, validated on 135 pages. It over-scores freshness and under-scores layout, neutrality and ownership.

It reads text, not the rendered page

Layout clarity and freshness are proxies: pop-ups vanish in extraction, and timestamps can reflect trivial edits.

Scope is 100 queries

Five intents; niche domains and other languages are untested. Mid-table rankings remain uncertain.

It is a snapshot

Systems, search results and pages keep changing. Overlap with Google is descriptive, not causal.

Takeaways

Source quality is a distinct axis from answer quality, and answer-centric benchmarks do not capture it.

- The top system's edge is source screening (ownership, authorship, domain authority) not relevance.
- It barely overlaps the search engine: only 16% of GPT-5's sources come from Google's first five pages.
- Retrieval quality and reasoning are partly interchangeable; a stronger retriever buys what reasoning buys.

Everything is public

Data: Harvard Dataverse, with Croissant and Responsible AI metadata: doi.org/10.7910/DVN/UOLXDQ

Code and the 100 benchmark queries: github.com/WukLab/SourceBench

Live leaderboard: huggingface.co/spaces/WukLab/SourceBench-Leaderboard