

NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

ListQA

A Benchmark for Evaluating List-Formatted Factual Knowledge Retrieval in Large Language Models

Mayank Kulkarni* · Rakesh Vaideeswaran*

Amazon AGI · *Equal contribution

Every frontier LLM has trained on Wikipedia. So we asked 34 of them list questions answered by Wikipedia's tables.

36.5%

best score, no thinking

<20%

most of the 34 models

LLM-Judge-F1 on ListQA, standard inference

THE TASK

Recall every row, filter across columns, return the complete list

SIMPLELISTQA · FLAT LIST · 59.5%

"Which songs by Bruno Mars were nominated for the MTV Video Music Award for Best Choreography between the years 2010 and 2024, where he was both the performer and the choreographer?"

1. The Lazy Song
2. Treasure
3. Finesse (Remix)

COMPLEXLISTQA · SUB-ATTRIBUTES · 40.5%

"In 2023 and 2024, for which awards was John Wick: Chapter 4 nominated in Best Stunts, who was the recipient and what was the result?"

1. Hollywood Critics Assoc. Midseason · Won
2. Las Vegas Film Critics Society · Nominated
3. St. Louis Film Critics Assoc. · Runner-up
4. Astra Film & Creative Arts Awards · Won

3–8 elements per answer · from parametric memory alone · nothing fabricated

THE DATA

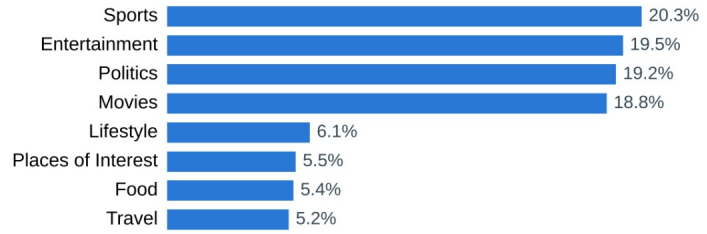
9,045 human-curated questions from 3,765 Wikipedia tables

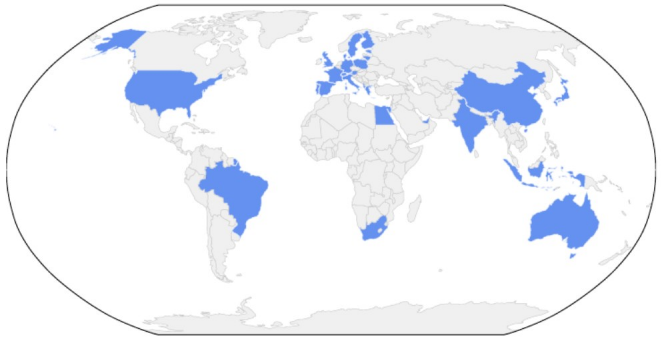
8
topics

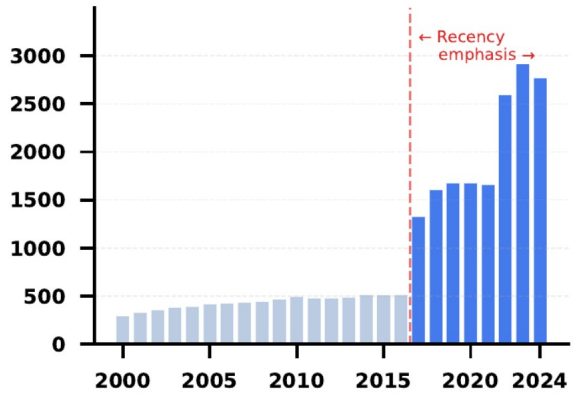
60+
countries

70%
fixed time window

1.7%
audited error rate



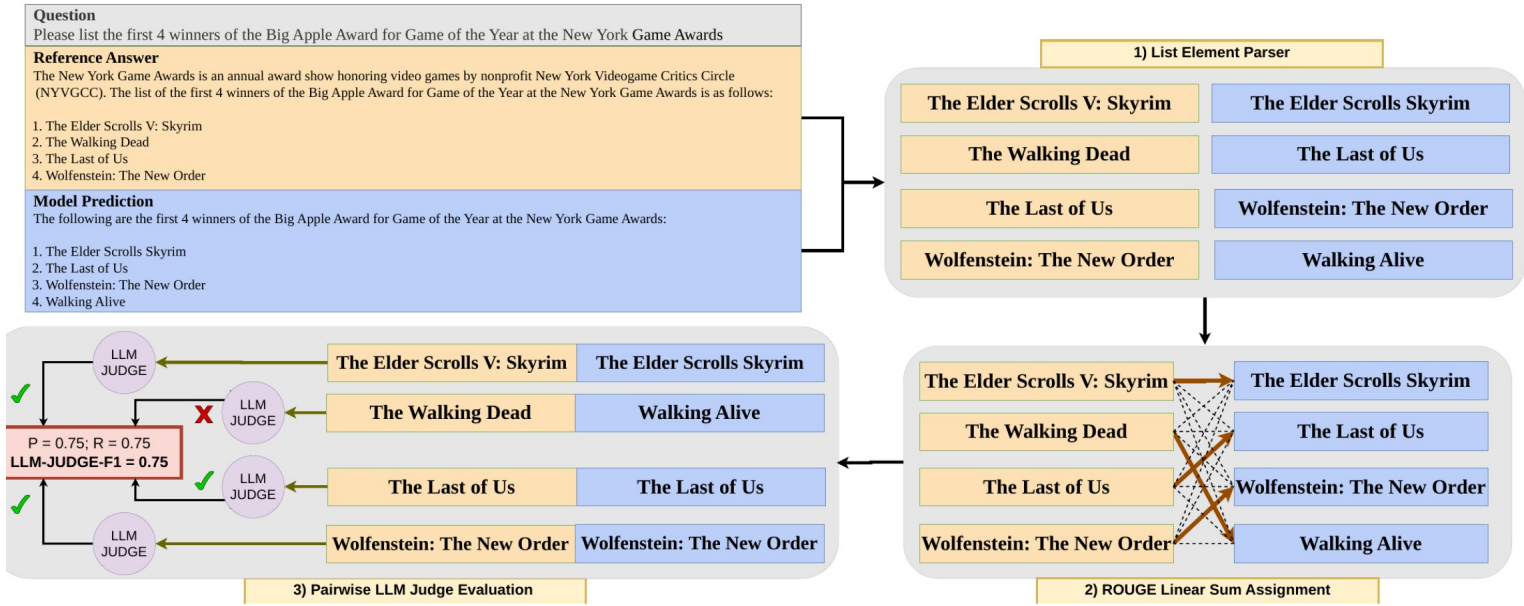




Evergreen: answers validated as of Dec 31, 2024 · 44 annotators, every pair cross-validated

SCORING

LLM-Judge-F1: match each item, then judge each pair



Catches hallucinations

Judging a whole list lets fabricated items slip through. Unmatched items count as wrong.

$O(n^2) \rightarrow O(n)$

Optimal one-to-one assignment means one judge call per element.

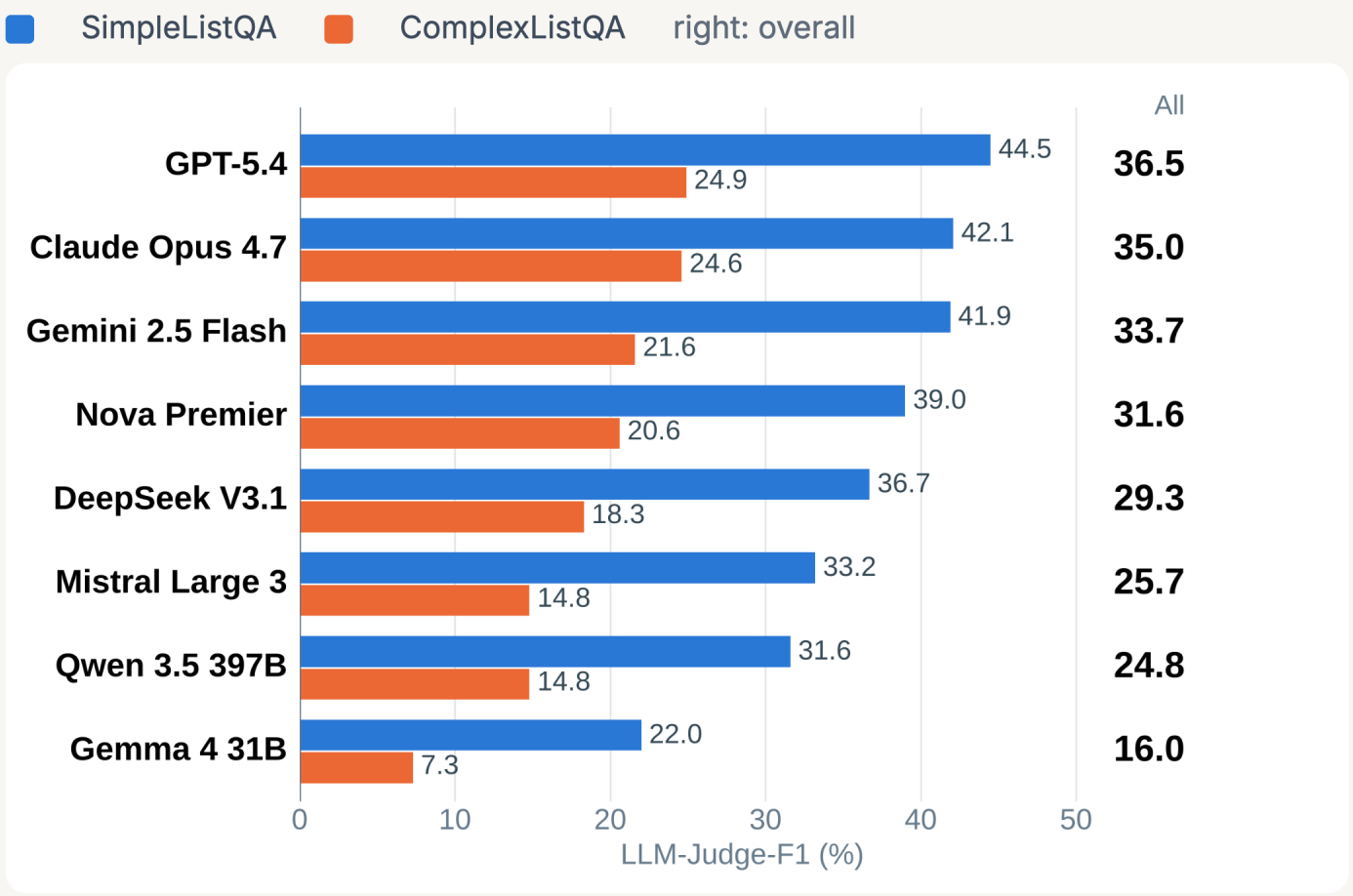
r = 0.965 with humans

A 4B judge ranks models like the 120B default (ρ = 0.972).

Paired with Format-Correctness to separate knowledge from instruction following

RESULTS · 34 MODELS, 8 FAMILIES

No model reaches 37% without thinking



44.5 → 24.9

GPT-5.4, flat vs hierarchical lists. Sub-attributes hurt every model.

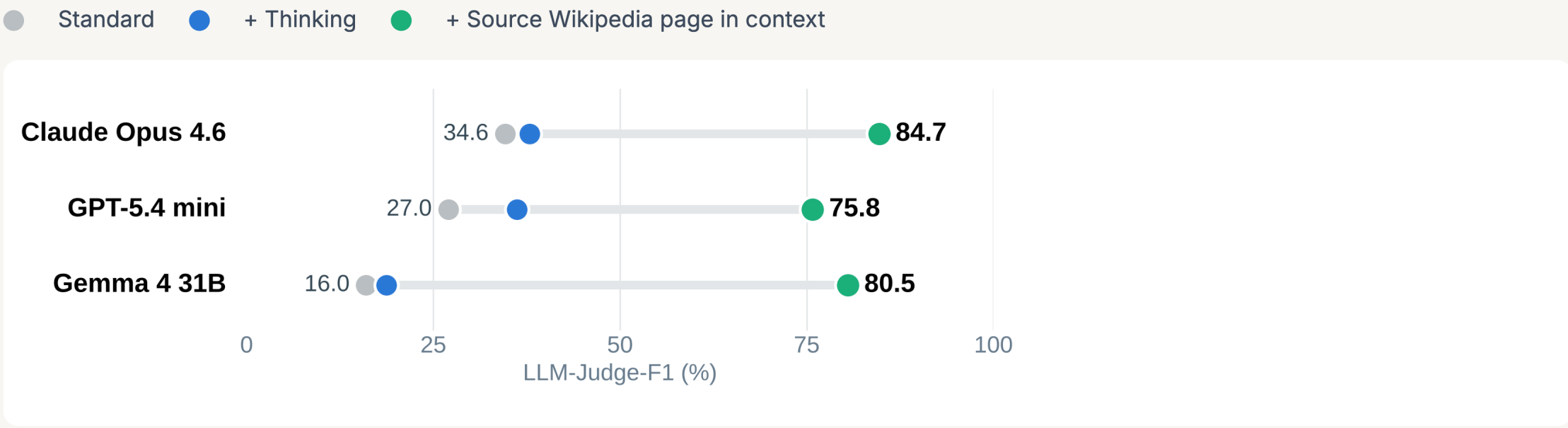
<20%

for most of the 34 models

LLM-Judge-F1 (%), best model per family, standard inference

WHAT MOVES THE SCORE

Knowledge is the bottleneck, not reasoning



+34 to +65 points
with the source page, for every model tested

Thinking helps less
Best: Gemini 3.1 Pro with thinking, 60.7

Recall tracks exposure
Every model: timeless > time-bounded; weakest on African and Asian entities

Oracle RAG and thinking results, LLM-Judge-F1 (%)

TAKEAWAYS

Recalling and composing structured knowledge from the weights is far from solved, even for facts every model has seen.

Science, law and medicine live in tables. ListQA's recipe works on any tabular source, to test whether training put the knowledge in the weights instead of paying for retrieval on every query.

listqa.github.io