INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2026

Microsoft Research

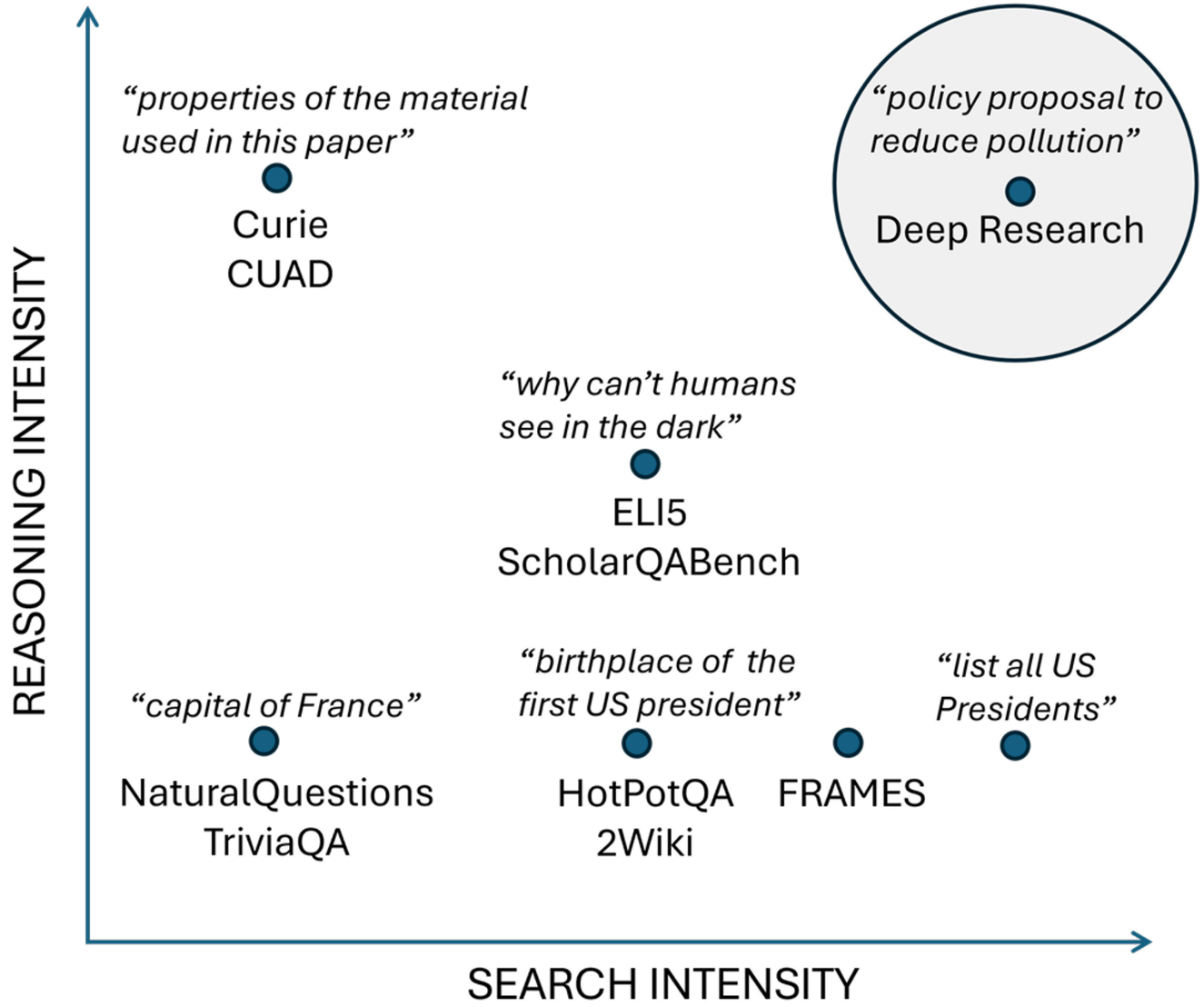
Characterizing Deep Research: A Benchmark and Formal Definition

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi,
Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta,
Nagarajan Natarajan, Amit Sharma

GITHUB.COM/MICROSOFT/LIVEDRBENCH

HUGGINGFACE.CO/DATASETS/MICROSOFT/LIVEDRBENCH

Background


REASONING INTENSITY
SEARCH INTENSITY
"properties of the material used in this paper"
Curie
CUAD
"policy proposal to reduce pollution"
Deep Research
"why can't humans see in the dark"
ELI5
ScholarQABench
"capital of France"
NaturalQuestions
TriviaQA
"birthplace of the first US president"
HotPotQA
2Wiki
FRAMES
"list all US Presidents"

Defining Deep Research (DR)

DEEP RESEARCH EXTENDS THE MULTI-HOP RETRIEVAL AUGMENTED GENERATION (RAG) TASK IN 3 ASPECTS

Search Intensity

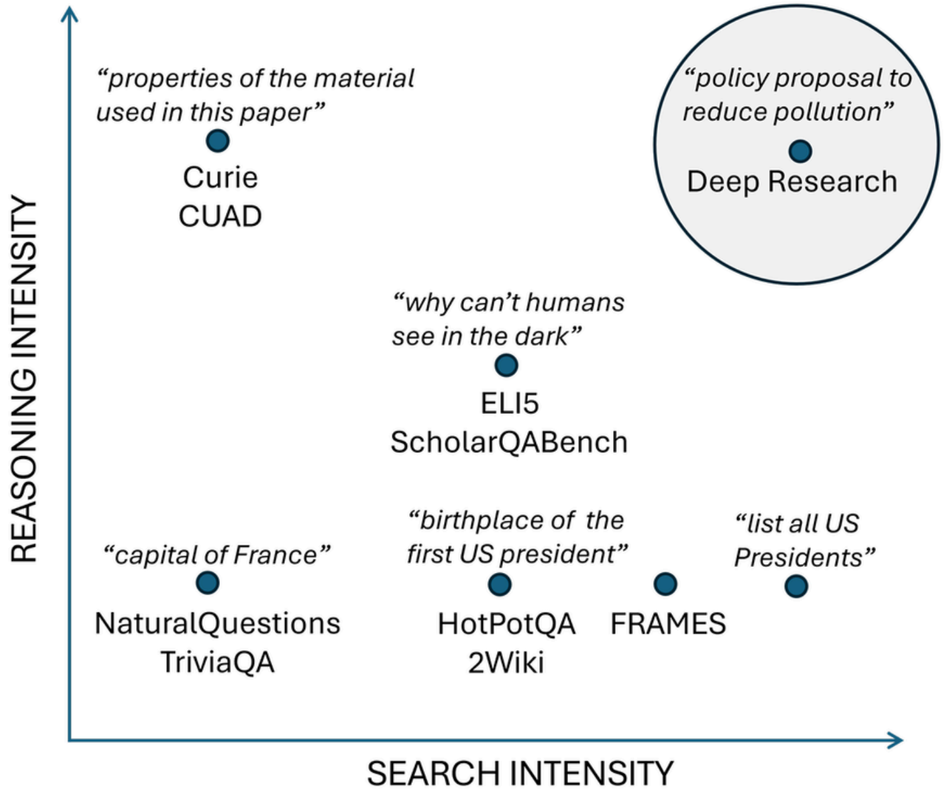
Processing of a large number of information units (paragraphs/ chunks); typically, ⩾20 information units via ⩾10 queries.

Reasoning Intensity

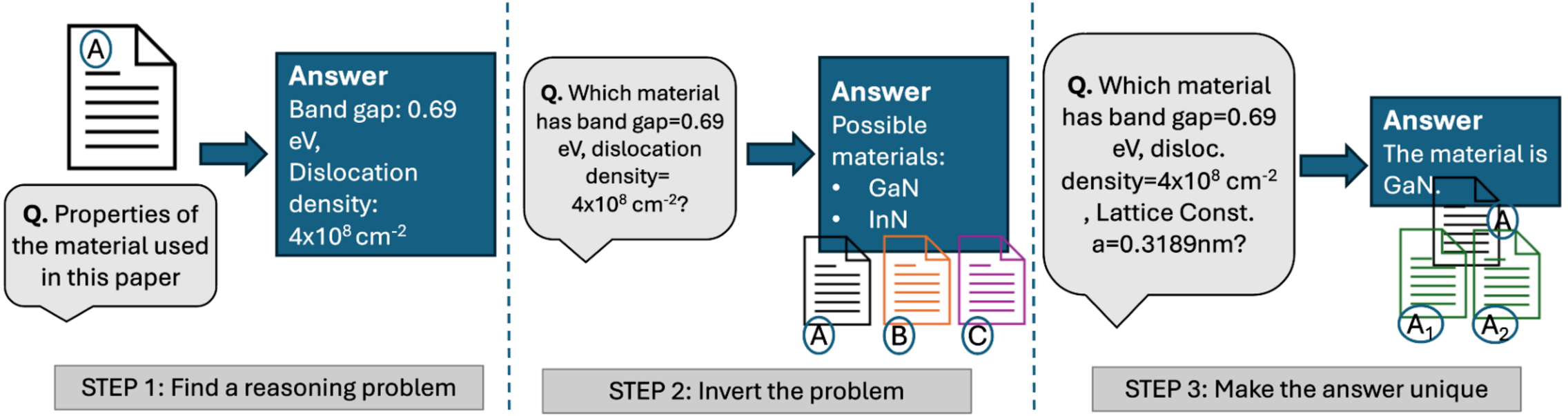
At least one subtask: finding, processing, or combining information, requires non-trivial reasoning.

Complexity of Outputs

The output of a deep research query is a list of claims, each of which may recursively contain sub-claims needed to derive it.



Creating a Benchmark using Problem Inversion



More questions can be created with new papers and public reports on the open web

LiveDRBench: 8 Task Categories

SciFacts Materials[1]

Find material matching measured properties

NovelDS Identification

Identify dataset matching unique characteristics and get its metadata.

NovelDS Peer Retrieval

Compare peer datasets in same problem space

NovelDS Extraction

Identify dataset + extract specific findings.

SciFacts Geo[1]

Find research papers that useall datasets specified in thequery.

Entities

Exhaustive list of real-world entities matching a criteria

Prior Art Search

Find any overlap between a proposed idea and existing literature.

Flights

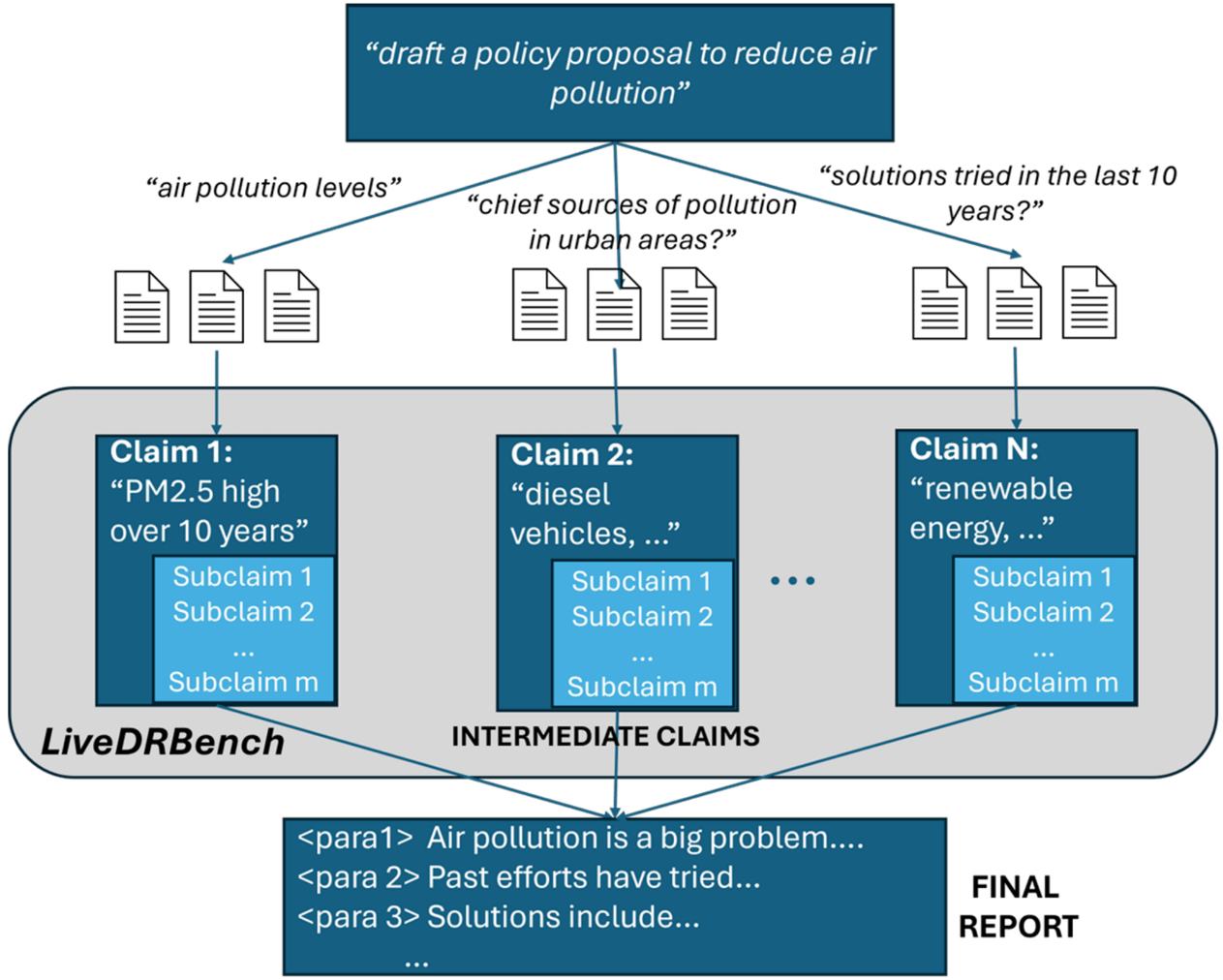
Identify flight incident from described conditions

[1] Cui et al., 2025, Curie: Evaluating llms on multitask scientific long-context understanding and reasoning

Benchmarking Deep Research

We focus not on the problem of writing a long report in DR,
but instead on the intermediate process of finding, processing, and combining information from many sources.

We evaluate this using precision and recall metrics over the generated claims.



Evaluation Setup

MODELS TESTED

DR Models

Gemini DR , Perplexity DR, OpenAI DR, DeepResearcher[2]

Reasoning Models

Gemini 2.5 Pro, Perplexity Sonar Reasoning, OpenAI o4-mini

Non-reasoning Models

Gemini 2.5 Flash, Perplexity's Sonar Pro, OpenAI GPT-4.1

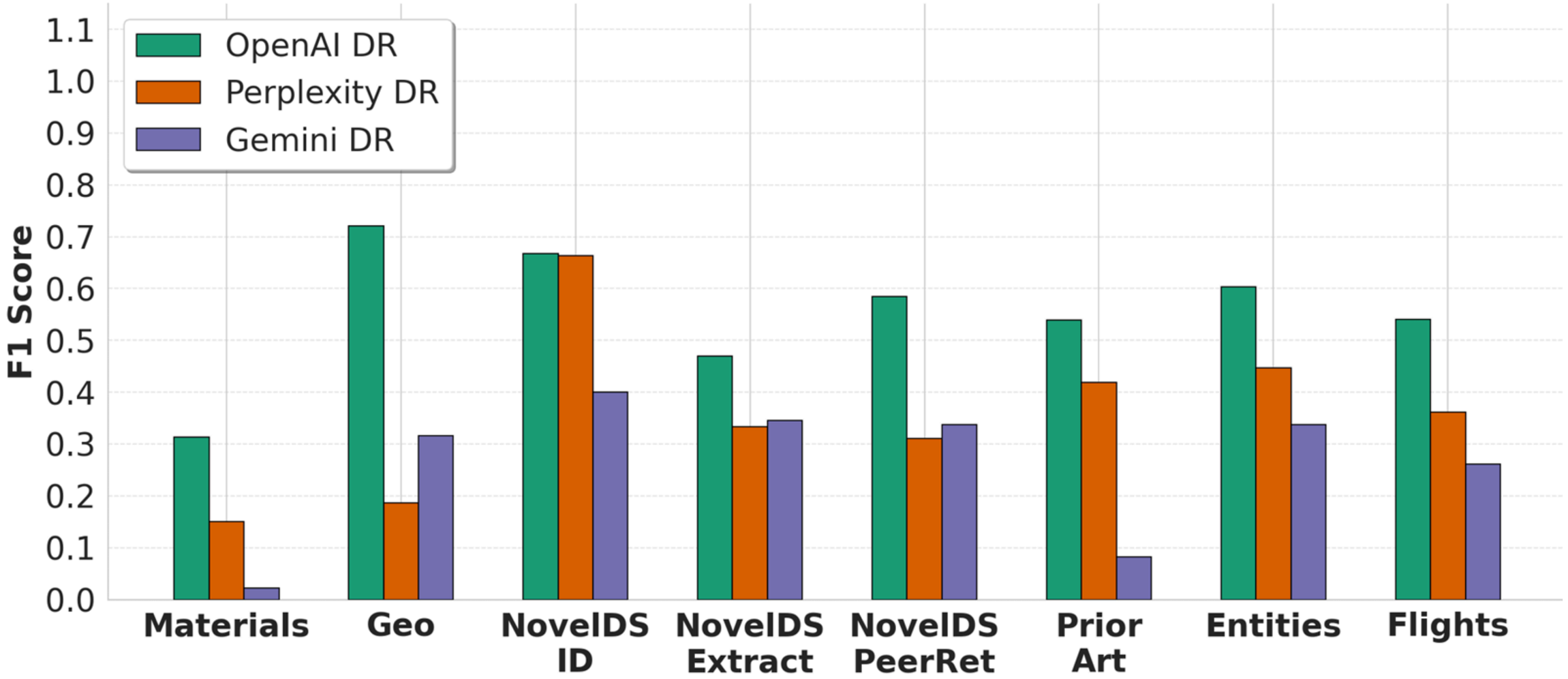
RECURSIVE CLAIM EVALUATION METRICS

$$\mathrm{Prec}(\mathcal{A}) = \frac{\sum_{A_i} w_i s(A_i)\,\mathrm{Prec}(A_i)}{\sum_{A_i} 1}$$

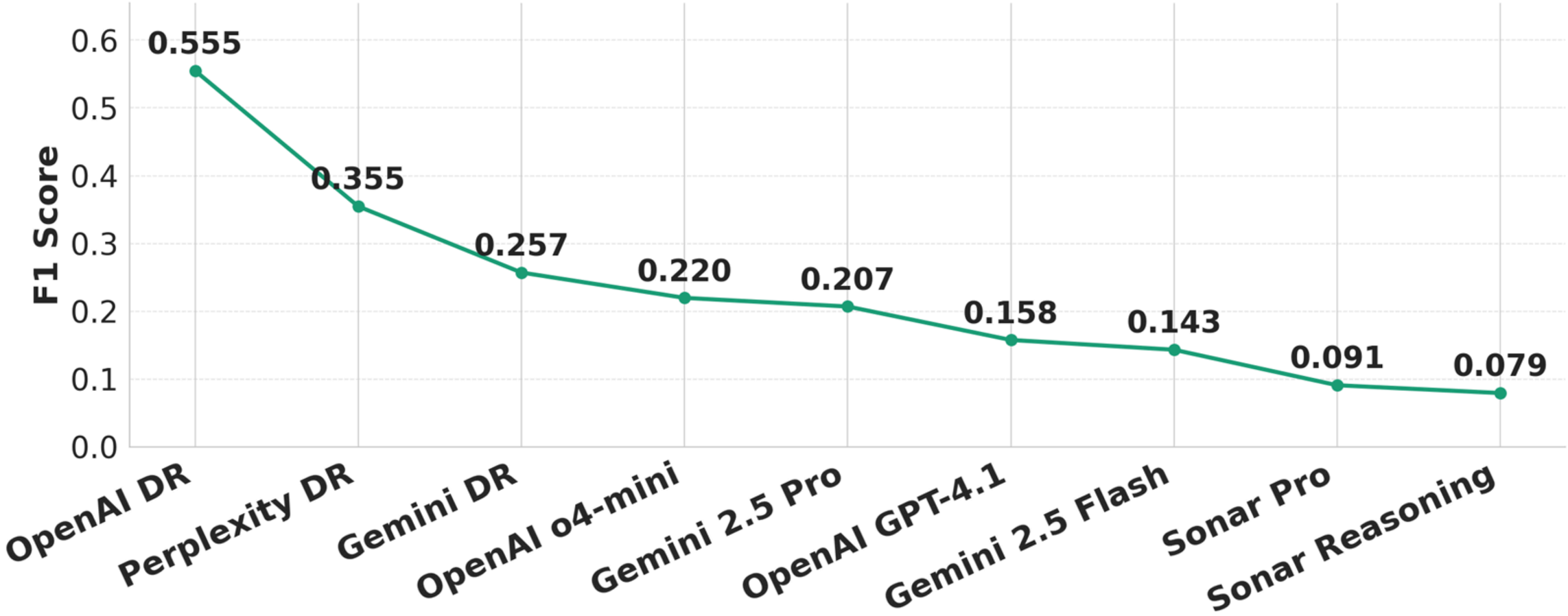
$$\mathrm{Rec}(\mathcal{A}) = \frac{\sum_{A_i} w_i s(A_i)\,\mathrm{Rec}(A_i)}{\sum_{A_i^*} 1}$$

[2] Zheng et al., 2025, Deepresearcher: Scaling deep research via reinforcement learning in real-world environments.

Overall Benchmark Results



Overall Benchmark Results



Search Behavior Analysis

Model	Necessary Query Coverage	Dependent Queries (Depth)	Independent Queries (Breadth)
OpenAI DR	**66.0%**	34	**64**
Perplexity DR	52.0%	15	25
Gemini DR	46.8%	**39**	24
DeepResearcher +DS Qwen 32B	53.4%	5	5

~50% Coverage Gap

DR Systems only cover about half the necessary search queries.

Proprietary Model Depth

Proprietary models issue 24–64 queries on average vs.5–6 for open-source DeepResearcher

Branching Matters

OpenAI DR has the highest branch count. Breadth ofexploration strongly correlates with F1

Necessary queries: Minimal set of queries required for discovery and extraction of the information asked in a DR query.

Conclusion

1. Formal definition: DR = high search intensity + high reasoning intensity + claims/sub-claims.
2. Introduce LiveDRBench, and claim based eval.
3. Problem inversion enables easy benchmark updates as the web evolves.
4. OpenAI DR leads at F1=0.55; large gap to open-source.
5. DR models cover only ~50% of necessary queries — search breadth is the bottleneck.

GITHUB.COM/MICROSOFT/LIVEDRBENCH

HUGGINGFACE.CO/DATASETS/MICROSOFT/LIVEDRBENCH