DATASETRESEARCH: BENCHMARKING LLM AGENTS FOR DEMAND-DRIVEN DATASET DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

As high-quality public data becomes scarce, sustaining model progress increasingly requires agents to discover datasets tailored to specific needs. Yet it remains unclear whether autonomous agents can reliably search for, transform, or synthesize useful training data. We introduce DATASETRESEARCH, comprising 208 reference-grounded demand specifications, each paired with a held-out reference dataset and detailed metadata. Our evaluation combines metadata alignment with downstream task performance under few-shot and fine-tuning protocols. Across open and proprietary systems, synthesis substantially outperforms search on Reasoning-intensive fine-tuning (75.46% vs. 34.52%), whereas Knowledge-intensive fine-tuning is close (43.67% vs. 45.29%) and the search advantage is clearer in few-shot evaluation. On the difficult 20-task `DatasetResearch-Pro` subset, a Retrieve–Transform–Synthesize workflow reaches 32.4% normalized fine-tune performance, compared with 24.1% for the strongest pure workflow; demand-only human reconstruction reaches 52.3%, demonstrating measurable headroom. These results establish DATASETRESEARCH as a discriminative benchmark for data-research agents while clarifying the scope and attainability of its reference-normalized scores.

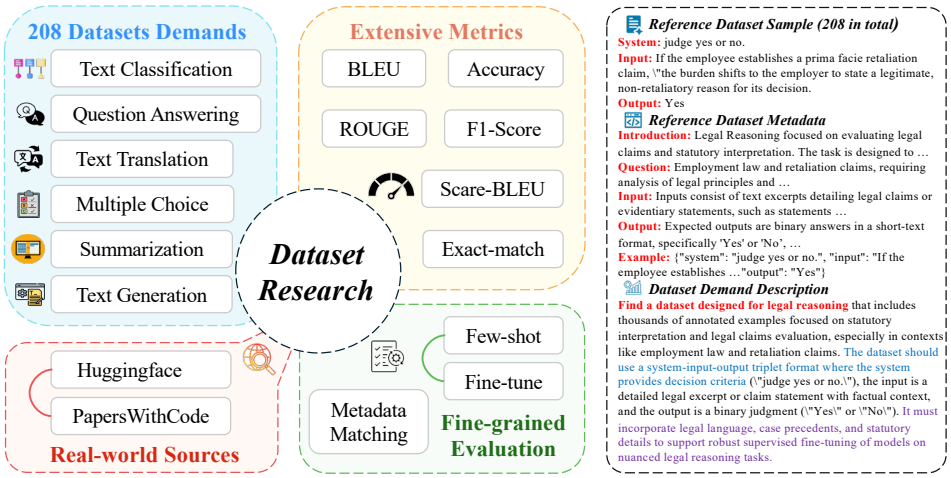

Figure 1: **Overview of DATASETRESEARCH.** The benchmark contains 208 reference-grounded demand specifications from two dataset sources, three evaluation dimensions, and six NLP metrics.

1 INTRODUCTION

The rapid evolution of AI (xAI, 2025a; OpenAI, 2025c; Anthropic, 2025; Comanici et al., 2025; Guo et al., 2025) has been largely fueled by the massive ingestion of public web data. However, as the field pushes the boundaries of scaling laws, a critical bottleneck has emerged: the exhaustion of high-quality, easily accessible public data (Xiao et al., 2025a; Li et al., 2026; Ye et al., 2025). The next frontier of model performance no longer relies on passive scraping (‘using what is available’),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but necessitates a paradigm shift toward active, demand-driven data discovery ('finding or creating what is needed'). This is particularly crucial for contemporary autonomous agentic systems, which exhibit a strong dependence on specialized, high-quality datasets to optimize searching and reasoning capabilities (Wang et al., 2025; Sun et al., 2024; Wang et al., 2024; Mai et al., 2025; Yang et al., 2025b; Xie et al., 2025).

Autonomous agents emerge as a promising vehicle for demand-driven dataset discovery, given the inherent complexity of translating abstract user needs into concrete data assets. Unlike passive scraping scripts, demand-driven dataset discovery requires iterative planning, searching, multi-step verification, and dynamic synthesis, which align precisely with recent advancements in agentic searching and reasoning (Google, 2025; OpenAI, 2025d; xAI, 2025b). Automating this rigorous pipeline is essential to overcome the scalability limits of manual curation, thereby improving the efficiency of the research lifecycle (Pan et al., 2025; OpenRouter, 2025; Kwa et al., 2025). However, the reliability of agents in demand-driven dataset discovery remains unproven. While general agent benchmarks exist (Wei et al., 2025; Xiao et al., 2025a;b), there is a lack of empirical frameworks tailored to the strict quality and functional requirements of dataset discovery. *Consequently, it is currently unknown whether autonomous agents can effectively search existing datasets or synthesize new datasets to meet specific dataset needs.*

To address this critical void, we introduce DATASETRESEARCH, the first comprehensive benchmark designed to rigorously evaluate autonomous agents that discover datasets from user needs. We curate 208 reference-grounded demand specifications and categorize them into *Knowledge-intensive* and *Reasoning-intensive* domains. These specifications are generated from reference metadata after removing the dataset identity and example; they are structured task descriptions rather than verbatim user logs. Each task is paired with a Reference Dataset and metadata covering its introduction, task, question, input, output, and sample information. Our DataResearcher framework searches for, transforms, or synthesizes dataset samples and converts them into standardized input–output pairs. Evaluation measures metadata alignment and downstream task performance through few-shot prompting and fine-tuning of LLaMA-3.1-8B (Dubey et al., 2024). An overview is shown in Figure 1.

We evaluate open-source, proprietary, and deep research systems, together with a Retrieve–Transform–Synthesize (RTS) hybrid. The full benchmark reveals an evaluation-dependent pattern: synthesis has a large advantage on Reasoning-intensive fine-tuning, while Knowledge-intensive fine-tuning is close and search has the clearer few-shot advantage. On the difficult 20-task `DatasetResearch-Pro` subset, RTS improves normalized fine-tune DTP from 24.1% for the best pure workflow to 32.4%. A retrieved-union oracle reaches 41.6% and demand-only human reconstruction reaches 52.3%, showing that the benchmark contains exploitable signal beyond current agents. A 30-task validity study further shows that concise natural-query rewrites preserve the system ranking ($\rho = 0.92$). Together, these diagnostics demonstrate both measurable headroom and robustness to query formulation without treating the reference score as a universally attainable ceiling.

In summary, our contributions are as follows:

(1) We formalize demand-driven dataset discovery and introduce DATASETRESEARCH, a benchmark of 208 reference-grounded demand specifications paired with comprehensive metadata and held-out reference datasets.

(2) We propose a systematic evaluation framework that moves beyond surface-level similarity, employing metadata alignment and downstream fine-tuning performance to measure the true utility of agent-discovered data.

(3) Through broad baselines, an RTS hybrid, attainable-headroom diagnostics, and a blinded query-validity study, we quantify current limitations while separating evaluation-dependent method differences from robust conclusions.

2 RELATED WORK

2.1 DATASET DISCOVERY

The critical role of high-quality datasets is well-recognized in recent research (Xiao et al., 2025a; Wu et al., 2025; Li et al., 2026; Xiao et al., 2025d; Fu et al., 2025; Xiao et al., 2025c;b), leading to

Table 1: **Composition of DATASETRESEARCH.** Reference Datasets in DATASETRESEARCH are divided into Knowledge-intensive and Reasoning-intensive. `DatasetResearch-Pro` is a subset containing 20 more challenging tasks.

Task Type	Metric	Knowledge 🤗	Knowledge [ ]	Reasoning 🤗	Reasoning [ ]	Number (Pro Subset)
Multiple Choice	Accuracy	9	4	4	10	5
Text Generation	BLEU	6	2	9	21	3
Text Summarization	ROUGE	2	1	0	8	3
Question Answering	F1, Exact Match	4	9	10	25	3
Text Classification	Accuracy	9	3	23	23	3
Language Translation	BLEU	1	1	14	10	3

two primary acquisition strategies: dataset searching and synthesis. Crucially, we distinguish this objective from Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023). While RAG retrieves specific context to augment model inference, dataset searching targets the upstream acquisition of comprehensive data assets for model training or fine-tuning. Consequently, searching often acts as a prerequisite step, identifying and curating the foundational knowledge bases that RAG systems subsequently utilize. Dataset searching methods include bi-encoder retrievers (Viswanathan et al., 2023; Soylu et al., 2024) and conversational AI (Walker et al., 2023; Majumder et al., 2024; Gu et al., 2024; Yang et al., 2025c), though the latter can hallucinate datasets. Dataset synthesis involves transforming existing data (Gandhi et al., 2024) or using agentic methods for construction, which has achieved strong results on benchmarks like SWE-bench verified (Jimenez et al., 2023; Yang et al., 2025b; 2024; Hui et al., 2024). The emergence of agents capable of iterative reasoning and searching (Zheng et al., 2025; Singh et al., 2025) necessitates a new standard for evaluation, motivating our development of DATASETRESEARCH.

2.2 LARGE LANGUAGE MODEL AGENTS

AI agent evolution is driven by stronger foundation models (xAI, 2025a; OpenAI, 2025c; Anthropic, 2025; Comanici et al., 2025; Guo et al., 2025; Yang et al., 2025a) with sophisticated searching and reasoning abilities (Zheng et al., 2025; Jin et al., 2025; Song et al., 2025; Lu et al., 2024). DATASETRESEARCH evaluates both types of agents. We assess search-enabled agents, including systems with integrated search tools (OpenAI, 2025a), GPT-5 (OpenAI, 2025b) and deep research systems (OpenAI, 2025d; Google, 2025; xAI, 2025b). For reasoning-based synthesis, we evaluate leading agents like DeepSeek-V3 Liu et al. (2024), OpenAI o3 (OpenAI, 2025c), GPT-5 OpenAI (2025b) and Claude-4.0-Sonnet (Anthropic, 2025). We use OpenAI o3's reasoning to construct challenging dataset generation tasks.

3 DATASETRESEARCH

In this section, we detail the construction of DATASETRESEARCH benchmark, the evaluated baseline agent DataResearcher and the comprehensive evaluation framework. DATASETRESEARCH composition is shown in Table 1, curation pipeline is shown in Figure 2, an overview of DataResearcher is shown in Figure 3 and DATASETRESEARCH evaluation pipeline is shown in Figure 4.

3.1 TASK DEFINITION

We formalize the problem of *Demand-Driven Dataset Discovery*, where an autonomous agent acts as a DataResearcher to curate a dataset matching specific user requirements. To rigorously evaluate this capability, we introduce *MetaTriplets* in DATASETRESEARCH. Formally, let $\mathcal{M} = \{M_i\}_{i=1}^{N=208}$ be a collection of datasets, where each dataset triplet is defined as $M_i = (D_i, S_{r_i}, M_{r_i})$ (1) **Demand Specification (D_i):** A reference-grounded natural language specification of the requested domain, task, schema, and desired features. It is generated from reference metadata after removing the dataset identity and example. (2) **Reference Dataset (S_{r_i}):** The ground-truth dataset sourced from real-world repositories that optimally satisfies D_i, serving as the gold standard for evaluation. (3) **Reference**

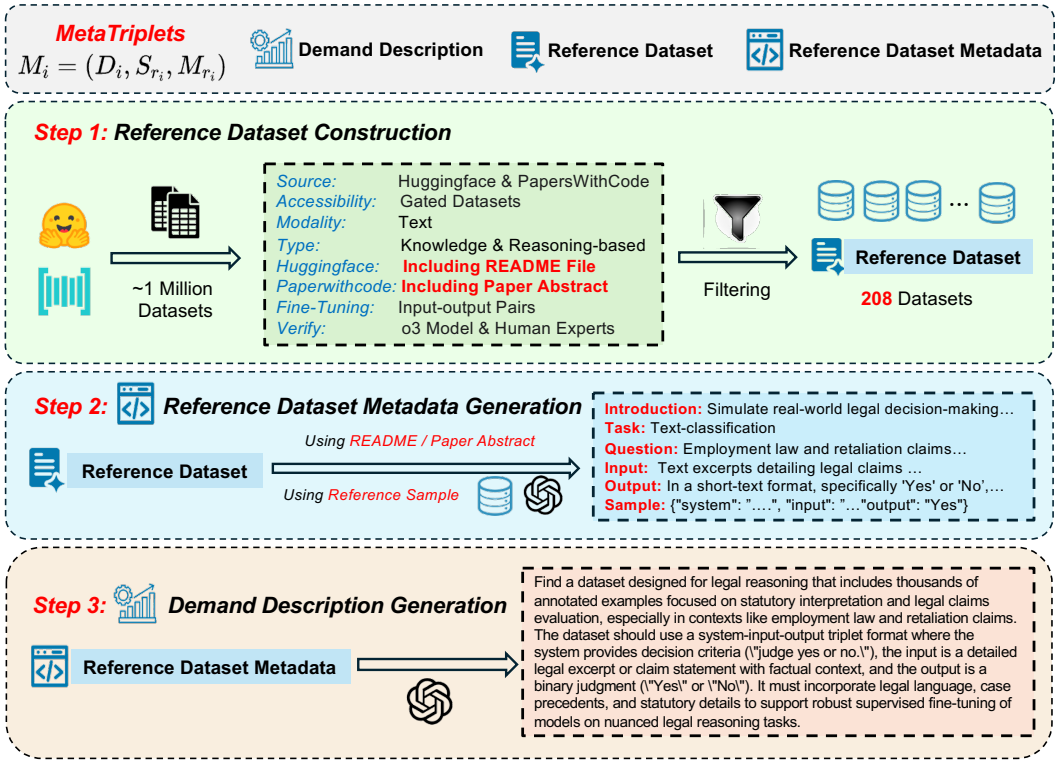

Figure 2: **Curation pipeline of DATASETRESEARCH.** We filter over one million candidates to obtain 208 Reference Datasets, use o3 to generate six-dimensional metadata, and convert that metadata into reference-grounded demand specifications after removing dataset identities and examples.

Dataset Metadata (M_{r_i}): A comprehensive profile of S_{r_i}, encapsulating detailed attributes including Reference Dataset introduction, task, question, input, output and a detailed sample.

Let $\mathcal{A}_\theta$ denote an agent parameterized by θ, capable of executing search or synthesis strategies. Given demand D_i, agent's goal is to generate a discovered dataset $S_d^{(i)} = \mathcal{A}_\theta(D_i)$. Core objective of DATASETRESEARCH is to quantify the semantic and functional gap between the agent-discovered set $S_d^{(i)}$ and the Reference Dataset S_{r_i}. We define the evaluation of agent-discovered dataset as an optimization problem minimizing the divergence $\mathbb{D}$ between the discovered and reference distributions:

$$\min_\theta \mathbb{E}_{M_i \sim \mathcal{M}} \left[\mathbb{D}(S_d^{(i)}, S_{r_i}) + \lambda \cdot \mathcal{L}_{\text{meta}}(S_d^{(i)}, M_{r_i}) \right] \tag{1}$$

where $\mathbb{D}(\cdot)$ represents the functional divergence measured via downstream task performance (i.e., how well a model trained on $S_d^{(i)}$ performs on the test distribution of S_{r_i}), and $\mathcal{L}_{\text{meta}}$ captures the semantic distance between the Discovered Dataset Metadata and the Reference Dataset Metadata. A lower total loss indicates a highly capable agent that can autonomously bridge the gap between abstract user demands and concrete, high-quality data assets.

3.2 BENCHMARK CURATION

DATASETRESEARCH comprises 208 reference-grounded demand specifications, with 91 reference datasets collected from HuggingFace and 117 from PaperswithCode. Each sample consists of a demand specification, a corresponding Reference Dataset and comprehensive Reference Metadata.

Data Collection Pipeline We implement a rigorous multi-stage pipeline designed to ensure data authenticity and evaluation feasibility in Figure 2. To distinguish discovery from direct retrieval, we

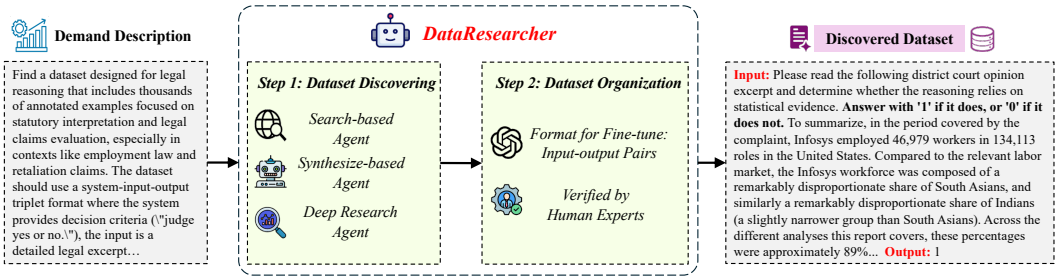

Figure 3: **Overview of the DataResearcher workflow.** Given a demand specification, DataResearcher searches for or synthesizes a dataset, generates its metadata, and compares it with the Reference Dataset Metadata through Figure 4.

initiate collection exclusively with 'Gated' datasets via the HuggingFace API; a 10-gram analysis finds less than 1% overlap between agent-discovered data and the gated references. We retain text datasets across six NLP categories (question answering, summarization, classification, generation, multiple choice, and translation) with appropriate metrics, sufficient documentation for metadata generation, and clear input–output mappings for fine-tuning. OpenAI o3 OpenAI (2025c) then converts the Reference Dataset Metadata into a demand specification while omitting the dataset identity and example (Section G). Human quality control covers the curation process. On 50 random instances, o3-generated and human-curated metadata achieve BERTScore > 0.95.

Knowledge and Reasoning intensive Tasks We further categorize DATASETRESEARCH tasks into two distinct types based on the learning objective. (1) **Knowledge-intensive** tasks (51 tasks) require extensive factual coverage and structured information, prioritizing the breadth and accuracy of the constructed data. (2) **Reasoning-intensive** tasks (157 tasks) emphasize logical deduction and pattern generalization, requiring data that guides the construction of reasoning pathways rather than relying on domain-specific memorization. During this process, four human experts independently classify the full dataset, achieving 100% inter-annotator agreement.

3.3 DATARESEARCHER

We evaluate four DataResearcher workflows. **Search-based Agents** query the HuggingFace Hub for up to five public datasets and select a valid candidate; gated references prevent trivial recovery of the target. **Deep Research Agents** conduct iterative web-scale discovery. **Synthesis-based Agents** generate 500 samples, either with one reference sample (*w/ ref*) or from the demand alone (*w/o ref*). The common pipeline is shown in Figure 3. Our additional **Retrieve–Transform–Synthesize (RTS)** workflow retrieves up to five packaged datasets or raw web sources, extracts and filters candidate examples, maps them to the requested schema, and synthesizes only missing coverage. RTS receives neither the reference identity or metadata nor a reference example. Detailed settings and prompts appear in Sections D and G.

Following discovery (searching or synthesis), DataResearcher utilize OpenAI o3 to parse all raw data into a standardized fine-tuning format with consistent input-output pairs, ensuring compatibility with our evaluation pipeline. Random testing of 50 samples shows that o3 achieved a 100% success rate in format conversion.

3.4 EVALUATION METHODOLOGY

We employ a multi-dimensional evaluation framework (Figure 4) assessing both intrinsic metadata alignment and extrinsic downstream task performance.

Metadata-Based Evaluation (Meta Score) We assess data quality by scoring the semantic alignment between the Reference Dataset Metadata and the discovered dataset's metadata. We employ OpenAI o3 model as a judge to assign scores (0-10) across six dimensions (introduction, task, question, input, output and sample). The final metadata alignment score (Meta Score) is calculated as the

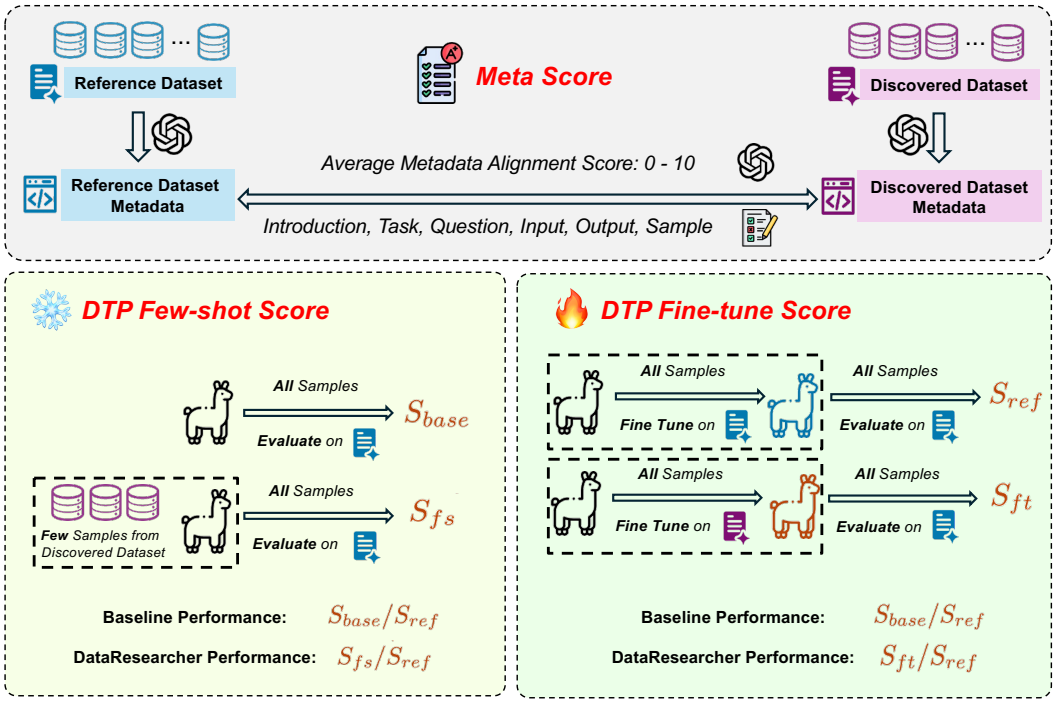

Figure 4: **DATASETRESEARCH Evaluation Pipeline.** We evaluate dataset discovery performance using LLaMA-3.1-8B on three metrics: Meta Score, DTP Few-shot Score and DTP Fine-tune Score. DTP is short for Downstream Task Performance.

average of these six dimensions, with a maximum score of 10. Since o3 correctly generates both the reference and discovered metadata verified by human experts, using it as the evaluator eliminates cross-model bias, ensuring a fair comparison within the same latent space. To further verify the objectivity of this metric, we conduct an ablation study using Claude-4-Sonnet as an independent judge. On a random subset of 50 samples, we observe a Pearson correlation coefficient > 0.95 with o3 scores, confirming that our metric reflects objective semantic alignment rather than model-specific stylistic preferences. Prompt is presented in Section G.

Downstream Task Performance (DTP Score) To quantify practical utility, we evaluate the performance of LLaMA-3.1-8B model (Dubey et al., 2024) on the Reference Dataset under three conditions: (1) **Zero-shot Baseline:** Direct evaluation of LLaMA-3.1-8B on Reference Dataset to establish a performance floor. (2) **Few-shot In-context Learning:** Few-shots using 1, 3, and 5 examples from the discovered dataset, then performing evaluation of LLaMA-3.1-8B on the Reference Dataset. (3) **Fine-tune:** Supervised fine-tuning of LLaMA-3.1-8B on the discovered dataset, then use the fine-tuned model to evaluate on the Reference Dataset. Detailed hyper-parameters and fine-tune configurations are provided in Section H.

Score Normalization Given the heterogeneity of metrics across the six task categories (e.g., BLEU for translation and Accuracy for classification), we implement a standardized normalization protocol. We define S_{ref} as the performance of LLaMA-3.1-8B fine-tuned directly on the Reference Dataset. This is a reference-normalized anchor, not a claim that every agent can attain it. For each evaluation setting, the score S_{eval} of a discovered dataset is normalized against this anchor. The normalized DTP Score is:

$$Score = \frac{S_{\text{eval}}}{S_{\text{ref}}} \times 100\% \tag{2}$$

This formulation scales performance relative to the reference condition. A score of 100% indicates parity under this protocol, while a score $> 100\%$ indicates higher measured performance than the reference condition.

Table 2: **Main Results of DATASETRESEARCH.** We report DTP 1-shot, DTP Fine-tune and Meta Score. Search-based GPT-5 and synthesis-based o3 are highlighted in gray. Subscripts denote the performance gap (red for improvement, blue for degradation).

DataResearcher	DTP Score (%): 1-shot		DTP Score (%): Fine-tune		Meta Score
	Knowledge↑	Reasoning↑	Knowledge↑	Reasoning↑	Average ↑
Baseline	$10.41_{-9.74}$	$11.84_{-4.00}$	–	–	–
Search-based					
GPT-4o-mini-Search	$6.48_{-13.67}$	$5.73_{-10.11}$	$12.12_{-33.17}$	$17.35_{-17.17}$	$5.52_{-1.74}$
GPT-4o-Search	$9.82_{-10.33}$	$7.25_{-8.59}$	$41.89_{-3.40}$	$27.54_{-6.98}$	$5.71_{-1.55}$
GPT-5	20.15	15.84	45.29	34.52	7.26
Synthesis-based					
Deepseek-V3	$8.97_{-1.28}$	$13.36_{-4.02}$	$28.16_{-10.82}$	$49.15_{-23.55}$	$8.25_{-0.44}$
Claude-4-Sonnet	$9.86_{-0.39}$	$15.89_{-1.49}$	$35.14_{-3.84}$	$65.16_{-7.54}$	$8.21_{-0.48}$
GPT-5	$12.44_{+2.19}$	$24.86_{+7.48}$	$43.67_{+4.69}$	$75.46_{+2.76}$	$8.97_{+0.28}$
OpenAI o3	10.25	17.38	38.98	72.70	8.69

4 EXPERIMENTS

We first compare search and synthesis across the full DATASETRESEARCH, then evaluate deep research, hybrid discovery, attainable headroom, and query-form sensitivity on the difficult 20-task `DatasetResearch-Pro` subset. We conclude with few-shot and retrieval ablations.

4.1 EXPERIMENTAL SETUP

On the full benchmark, search-based agents use GPT-4o-mini-Search-Preview, GPT-4o-Search-Preview (OpenAI, 2025a), and GPT-5 (OpenAI, 2025b); synthesis-based agents use DeepSeek-V3 (Liu et al., 2024), Claude-4-Sonnet (Anthropic, 2025), OpenAI o3 (OpenAI, 2025c), and GPT-5. On `DatasetResearch-Pro`, we additionally evaluate OpenAI, Gemini, and Grok Deep Research (OpenAI, 2025d; Google, 2025; xAI, 2025b) and RTS. All systems in the additional Pro-20 comparison share a 500-example output budget. The oracle uses held-out DTP only as a non-deployable selector over the retrieved union. Human researchers receive the demand and web access but no reference identity, metadata, or examples.

4.2 MAIN RESULTS

Table 2 reports the primary results (expanded in Table 5). The paradigm comparison depends on both task type and evaluation. For Reasoning-intensive fine-tuning, GPT-5 synthesis reaches 75.46%, compared with 34.52% for GPT-5 search. For Knowledge-intensive fine-tuning, however, the corresponding scores are 43.67% and 45.29%, a difference of only 1.62 points; the search advantage is clearer in one-shot evaluation (20.15% vs. 12.44%). GPT-5 synthesis also exceeds OpenAI o3 across the reported settings, including a 7.48-point gain on one-shot Reasoning. These results support

Table 3: **Single Reference Sample Guidance Influence.** Experiments with o3 indicate that while providing a reference sample improves the final score, the overall gain is limited.

Setting	DTP Score: Fine-tune		Meta Score
	Knowledge(%)↑	Reasoning(%)↑	Average ↑
w/o ref	37.94	67.25	8.57
w/ ref	$38.98_{+1.04}$	$72.70_{+5.45}$	$8.69_{+0.12}$

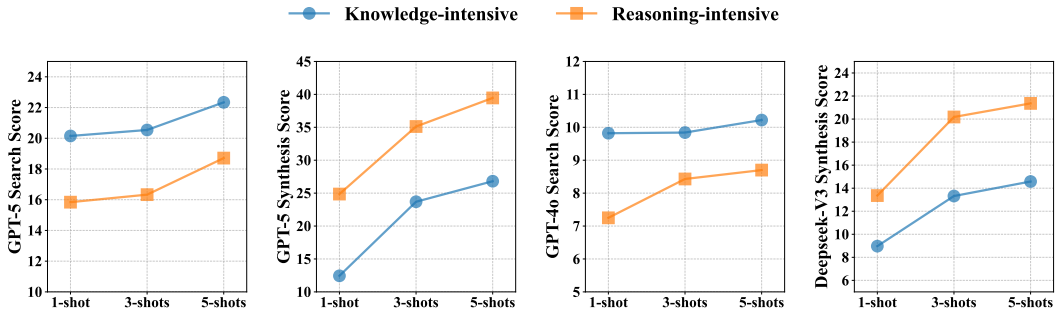

Figure 5: **Comparison between Search-based (Left) and Synthesis-based (Right) DataResearcher across Few-shots Settings.** Search-based agents show superior performance in Knowledge-intensive tasks, whereas synthesis-based models are more effective for Reasoning-intensive tasks.

a strong synthesis advantage for Reasoning-intensive fine-tuning, but not a blanket search–synthesis ordering.

Synthesis-based agents attain higher metadata alignment (up to 8.97 for GPT-5), reflecting closer adherence to the requested schema than retrieved repositories. This benefit comes with greater generation cost, motivating workflows that reuse external data when possible.

4.3 Few-shot Scaling Dynamics

Figure 5 shows 1-, 3-, and 5-shot performance. Search-based models begin strongly on Knowledge-intensive tasks but gain little from additional demonstrations, consistent with fixed retrieved content. Synthesis-based models improve more clearly with additional shots on Reasoning-intensive tasks. Across both paradigms, three shots provide the most consistent cost–performance trade-off.

4.4 Hybrid Workflows and Attainable Headroom

Figure 6 summarizes the originally reported pure and Deep Research systems. Under the matched additional `DatasetResearch-Pro` protocol, GPT-5 Search, GPT-5 Synthesis, and OpenAI Deep Research obtain normalized fine-tune DTP scores of 18.8%, 24.1%, and 22.1%, respectively. RTS instead retrieves, transforms, and selectively fills missing coverage, reaching 32.4%. This is 8.3 points above the strongest pure workflow and 10.3 points above Deep Research, showing that the benchmark rewards a materially stronger data-research strategy rather than uniformly assigning low scores. Two diagnostics establish attainable headroom. A best-of-retrieved-union oracle reaches 41.6%, while demand-only human reconstruction reaches 52.3%; in 18 of 20 tasks, the human score exceeds the best autonomous score by at least five points. Fine-tuning on 500 reference examples reaches 94.2%, and using up to 1,000 reaches the 100.0% normalization anchor. Thus, sample count alone does not explain current agent performance, while the remaining human–reference gap also cautions against interpreting 100% as a universally attainable agent ceiling.

Table 4: **Comparison against Open Retrieval Stack Baseline.** Simple open retrieval stack is insufficient for dataset discovery.

Setting	DTP Score: 1-shot		Meta Score
	Knowledge(%)↑	Reasoning(%)↑	Average ↑
Retrieval Baseline	5.38	4.13	4.56
GPT-4o-Search	9.82$_{+4.44}$	7.25$_{+3.12}$	5.71$_{+1.15}$

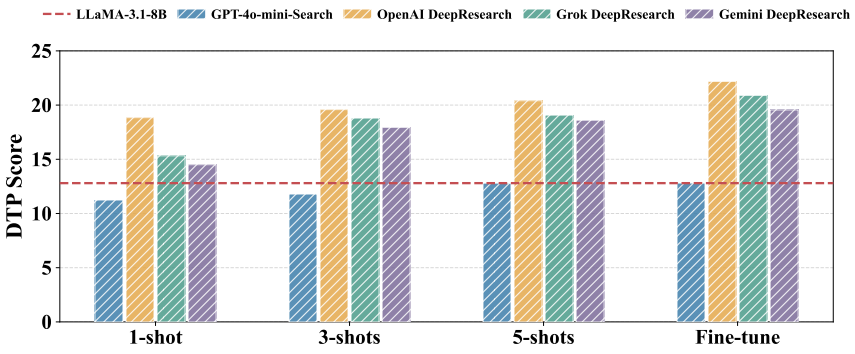

Figure 6: **Pure-workflow DTP on `DatasetResearch-Pro`,** the 20 tasks where GPT-4o-mini-Search performs worst. Plotted workflows remain below 25%; RTS is reported in the text. The dashed line is the LLaMA-3.1-8B reference baseline.

4.5 Demand Validity and Query-form Sensitivity

Benchmark inputs are reference-grounded specifications, not verbatim query logs. We therefore evaluate ecological validity directly. Ten data-experienced researchers supplied 30 anonymized natural-query rewrites, and three blinded raters compared them with matched original specifications. Original and rewritten forms were judged usable for a real project in 82% and 88% of cases, with actionability scores of 4.2 and 4.3 out of 5. Source-identification accuracy decreases from 66% to 54% under the concise rewrites, yet agent rankings remain highly stable ($\rho = 0.92$) and mean absolute DTP change is 2.3 points. We therefore report specification and natural-query tracks separately: the former is intentionally explicit, while the latter tests sensitivity to more concise requests.

4.6 Ablation Studies

We isolate two additional factors. (1) **Reference guidance:** comparing OpenAI o3 without and with one reference sample (Table 3) yields gains of 1.04 points on Knowledge and 5.45 points on Reasoning; substantial zero-shot performance shows that synthesis is not merely copying the example. (2) **Open retrieval stack:** BM25 (Robertson et al., 2009) with ColBERT (Khattab & Zaharia, 2020) obtains 5.38% on Knowledge and 4.13% on Reasoning (Table 4), far below GPT-4o-Search DataResearcher. Demand-driven discovery therefore requires semantic alignment beyond keyword overlap.

5 Discussion

Conclusion In this work, we formalized the paradigm of demand-driven dataset discovery and introduced DatasetResearch, a rigorous benchmark designed to evaluate the readiness of autonomous agents for this critical task. By subjecting state-of-the-art systems to 208 real-world requirements, our evaluation reveals that while agents demonstrate promise in specialized sub-domains—with search-based methods excelling in knowledge retrieval and synthesis-based approaches in reasoning—they collectively struggle to achieve true autonomy, particularly on out-of-distribution tasks. These findings underscore the necessity of DatasetResearch as a standard for measuring progress toward agents capable of sustaining scaling laws through active data engineering.

Limitations and Future Work Our current investigation concentrates on the textual domain, establishing a foundational protocol for evaluating dataset discovery where the scarcity of high-quality data is most immediately felt. As the broader field evolves toward omni-modal capabilities, we recognize that the scope of autonomous data curation must expand beyond text. Future iterations of this work aim to incorporate multimodal data requirements, such as vision-language pairs and audio sequences, thereby extending our evaluation framework to support the development of next-generation multimodal autonomous agents.

REFERENCES

Anthropic. Claude-4, 2025. URL https://www.anthropic.com/news/claude-4.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xuan-Quy Dao, Ngoc-Bich Le, The-Duy Vo, Xuan-Dung Phan, Bac-Bien Ngo, Van-Tien Nguyen, Thi-My-Thanh Nguyen, and Hong-Phuoc Nguyen. Vnhsge: Vietnamese high school graduation examination dataset for large language models, 2023. URL https://arxiv.org/abs/2305.12199.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Dayuan Fu, Yunze Wu, Xiaojie Cai, Lyumanshan Ye, Shijie Xia, Zhen Huang, Weiye Si, Tianze Xu, Jie Sun, Keyu Li, et al. Interaction as intelligence part ii: Asynchronous human-agent rollout for long-horizon task training. *arXiv preprint arXiv:2510.27630*, 2025.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. Better synthetic data by retrieving and transforming existing datasets. *arXiv preprint arXiv:2404.14361*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Google. Gemini-deepresearch, 2025. URL https://gemini.google/overview/deep-research/?hl=en.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. Blade: Benchmarking language model agents for data-driven science. *arXiv preprint arXiv:2408.09667*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

William JF Keenan. Sacre bleu: Faith, fashion and freedom: Marist foundation garments 1817–1862. In *Materializing Religion*, pp. 132–153. Routledge, 2017.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, et al. Measuring ai ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Shijie Xia, Xiaojie Cai, Tianze Xu, Weiye Si, et al. Agencybench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts. *arXiv preprint arXiv:2601.11044*, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2024.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

Xinji Mai, Haotian Xu, Weinong Wang, Yingying Zhang, Wenqiang Zhang, et al. Agent rl scaling law: Agent rl with spontaneous code execution for mathematical problem solving. *arXiv preprint arXiv:2505.07773*, 2025.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

OpenAI. Gpt-4o-search-preview, 2025a. URL `https://platform.openai.com/docs/models/gpt-4o-search-preview`.

OpenAI. Gpt-5. `https://openai.com/zh-Hans-CN/gpt-5/`, 2025b.

OpenAI. Openai o3, 2025c. URL `https://openai.com/zh-Hans-CN/index/introducing-o3-and-o4-mini/`.

OpenAI. Openai-deepresearch, 2025d. URL `https://openai.com/zh-Hans-CN/index/introducing-deep-research/`.

OpenRouter. State of ai: An empirical 100 trillion token study with openrouter. `https://openrouter.ai/state-of-ai`, 2025.

Melissa Z Pan, Negar Arabzadeh, Riccardo Cogo, Yuxuan Zhu, Alexander Xiong, Lakshya A Agrawal, Huanzhi Mao, Emma Shen, Sid Pallerla, Liana Patel, et al. Measuring agents in production. *arXiv preprint arXiv:2512.04123*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*, 2019.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 3(4):333–389, 2009.

Ofir Ben Shoham and Nadav Rappoport. Medconceptsqa: Open source medical concepts qa benchmark. *Computers in Biology and Medicine*, 182:109089, 2024.

Ramneet Singh, Sathvik Joel, Abhav Mehrotra, Nalin Wadhwa, Ramakrishna B Bairi, Aditya Kanade, and Nagarajan Natarajan. Code researcher: Deep research agent for large systems code and commit history. *arXiv preprint arXiv:2506.11060*, 2025.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.

Dilara Soylu, Christopher Potts, and Omar Khattab. Fine-tuning and prompt optimization: Two great steps that work better together. *arXiv preprint arXiv:2407.10930*, 2024.

Chuanneng Sun, Songjun Huang, and Dario Pompili. Llm-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*, 2024.

Vijay Viswanathan, Luyu Gao, Tongshuang Wu, Pengfei Liu, and Graham Neubig. Datafinder: Scientific dataset recommendation from natural language descriptions. *arXiv preprint arXiv:2305.16636*, 2023.

Johanna Walker, Elisavet Koutsiana, Joe Massey, Gefion Thuermer, and Elena Simperl. Prompting datasets: Data discovery with conversational agents. *arXiv preprint arXiv:2312.09947*, 2023.

Xiangwen Wang, Jie Peng, Kaidi Xu, Huaxiu Yao, and Tianlong Chen. Reinforcement learning-driven llm agent for automated attacks on llms. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pp. 170–177, 2024.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Xiangkun Hu, et al. Innovatorbench: Evaluating agents' ability to conduct innovative llm research. *arXiv preprint arXiv:2510.27598*, 2025.

xAI. Grok-4, 2025a. URL `https://x.ai/news/grok-4`.

xAI. Grok-deepresearch, 2025b. URL `https://x.ai/news/grok-3`.

Yang Xiao, Mohan Jiang, Jie Sun, Keyu Li, Jifan Lin, Yumin Zhuang, Ji Zeng, Shijie Xia, Qishuo Hua, Xuefeng Li, et al. Limi: Less is more for agency. *arXiv preprint arXiv:2509.17567*, 2025a.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. Towards dynamic theory of mind: Evaluating llm adaptation to temporal evolution of human states. *arXiv preprint arXiv:2505.17663*, 2025b.

Yang Xiao, Jiashuo Wang, Ruifeng Yuan, Chunpu Xu, Kaishuai Xu, Wenjie Li, and Pengfei Liu. Limopro: Reasoning refinement for efficient and effective test-time scaling. *arXiv preprint arXiv:2505.19187*, 2025c.

Yang Xiao, Chunpu Xu, Ruifeng Yuan, Jiashuo Wang, Wenjie Li, and Pengfei Liu. Scale: Selective resource allocation for overcoming performance bottlenecks in mathematical test-time scaling. *arXiv preprint arXiv:2512.00466*, 2025d.

Jingxu Xie, Dylan Xu, Xuandong Zhao, and Dawn Song. Agentsynth: Scalable task generation for generalist computer-use agents. *arXiv preprint arXiv:2506.14205*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

John Yang, Kilian Leret, Carlos E Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. Swe-smith: Scaling data for software engineering agents. *arXiv preprint arXiv:2504.21798*, 2025b.

Wenli Yang, Rui Fu, Muhammad Bilal Amin, and Byeong Kang. The impact of modern ai in metadata management. *Human-Centric Intelligent Systems*, pp. 1–28, 2025c.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

A AI USE STATEMENT

We leveraged LLMs to generate the metadata portion of the dataset, as well as to support LLM-as-a-judge evaluation, with the generated content undergoing rigorous human inspection and cross-model validation. We also used LLMs to polish the writing of our paper.

B PRACTICAL USE CASES FOR DATASET DISCOVERY

To further justify the motivation behind DATASETRESEARCH, we present a concrete user scenario illustrating the necessity of autonomous demand-driven dataset discovery where generic datasets fall short.

Scenario: Specialized Legal Model Training Consider a legal researcher or practitioner aiming to fine-tune a small language model (SLM) specifically for predicting judgments in *'Labor Law Retaliation Claims.'*

- **Status Quo (Manual Workflow):** The user faces significant barriers. Generic legal datasets (e.g., Pile-of-Law) are often too broad and lack the specific case law logic required for this niche. The user must spend days manually searching GitHub repositories, legal forums, or government databases. Alternatively, they must write complex, rule-based scripts to scrape and filter data, requiring significant engineering expertise.
- **The Agentic Solution (DataResearcher):** Using our proposed framework, the user simply inputs a natural language demand: *'Find or synthesize a dataset of labor law retaliation claims containing case facts and final judgments.'*
  - **Discovery Mode:** The agent autonomously scans the web to locate niche, pre-existing datasets that a human keyword search might miss.
  - **Synthesis Mode:** If no direct dataset exists, the agent analyzes the logic of relevant case law and synthesizes high-quality, privacy-preserving training samples that mimic the structure of real judgments.

Value Proposition: This capability lowers the barrier for Domain-Specific AI development, allowing domain experts (lawyers, doctors, scientists) to create high-quality training assets without deep data engineering skills.

C EXTENDED ANALYSIS: PARADIGMS AND BASELINES

C.1 DIFFERENTIATION FROM RAG AND STATIC SYNTHESIS

A critical distinction exists between our demand-driven discovery agent and existing methodologies such as Retrieval-Augmented Generation (RAG) or task-specific synthesis pipelines (e.g., WizardMath Luo et al. (2023)).

Discovery vs. RAG RAG is primarily an *inference-time* technique that retrieves context to answer queries. In contrast, Dataset Discovery is a *training-side* process aimed at acquiring data assets to fine-tune models. As demonstrated in our benchmark, we evaluate the utility of the discovered data by fine-tuning a LLaMA-3.1-8B model. Furthermore, dataset discovery is often the prerequisite step for building high-quality RAG systems (i.e., finding the domain-specific documents to populate the RAG knowledge base).

Discovery vs. Static Pipelines Works like WizardMath utilize hard-coded, static pipelines tailored for a single domain. DATASETRESEARCH evaluates an agent's ability to *autonomously plan* and execute strategies for *any* open-ended user demand. The core challenge is adaptability: can the agent automatically reproduce a WizardMath-like effect for a math prompt, and purely switch to a retrieval-heavy strategy for a historical query? Our findings regarding the search-synthesis dichotomy provide the heuristic foundation for such adaptive agents.

D EVALUATION SETUP

We evaluate leading agents on the 208-task DATASETRESEARCH and the 20-task `DatasetResearch-Pro`. Systems include search-based, synthesis-based, deep research, and RTS workflows, with oracle and human conditions used only as validity diagnostics. Evaluation is twofold: (1) Metadata-based Evaluation reflects instruction following, and (2) Downstream Task Performance (DTP) measures the functional utility of the discovered set. We calculate six metrics tailored to different task categories:

- **Accuracy** measures the proportion of correct predictions, reflecting models' ability to identify the correct option or label.
- **F1-Score** (Joshi et al., 2017) computes the harmonic mean of precision and recall, providing a balanced assessment of model performance, especially when dealing with partial matches or token-level evaluation.
- **Exact Match** (Joshi et al., 2017) evaluates the percentage of predictions that exactly match the ground-truth answers, offering a strict assessment criterion.
- **BLEU** (Papineni et al., 2002) measures n-gram overlap between generated and reference text, assessing the quality and fluency of generated content.
- **SacreBLEU** (Keenan, 2017) provides a standardized and reproducible version of BLEU scoring, ensuring consistent evaluation.
- **ROUGE** (Lin, 2004) calculates recall-oriented overlap of n-grams and longest common subsequences, specifically designed for evaluating text summarization quality and content preservation.

DataResearcher Settings To ensure a fair comparison, the Data Researchers have the following experimental settings:

- Search agents evaluated on DATASETRESEARCH are restricted to public datasets on HuggingFace to facilitate automated data formatting and processing. For `DatasetResearch-Pro`, search scope is expanded to the entire web, with results manually curated for relevance and accessibility.
- For all datasets returned by search agents, we use the OpenAI o3 to parse the data into a standardized fine-tuning format and to generate their corresponding metadata, mirroring the format of our reference set. If a discovered dataset contained more than 1000 samples, we truncate it to the first 1000 sample pairs.
- For synthesis agents, we generate datasets of 500 samples by prompting the model to produce 10 samples at a time and concatenating the results over 50 iterations. This mitigates potential quality degradation from long context windows.
- Deep Research Agents employ advanced browsing capabilities from OpenAI, Grok, and Gemini to perform web-scale discovery. To evaluate these closed deep research systems without API access, we enforce a strict Human-in-the-loop protocol: the agent executes the search and identifies the data source (providing URLs or generation code), while a human operator performs the mechanical retrieval and format conversion. Crucially, the human operator acts solely as an executor and does not perform data filtering or cleaning.

E METADATA

Our benchmark provides comprehensive metadata annotations for each task instance to facilitate systematic evaluation and analysis. Each MetaTriplet contains the following structured metadata components:

- **Introduction**: The task and area description of this task instance.
- **Task**: The classification of task type.
- **Question**: Question Content Type - Describes the primary knowledge domains and content types covered by the questions in the test dataset.

- **Input**: Structured retrieval results and contextual information - Input consists of formatted search results containing metadata fields such as descriptions, display URLs, titles, actual URLs, and ranking information, along with potential tabular data, document snippets, and conversational dialogue history for multi-turn scenarios.
- **Output**: Direct factual answer format - Outputs are concise, definitive answers that directly address the question based on the provided context, formatted as complete statements such as 'The answer is [specific fact]' for factual queries, numerical values for arithmetic problems, and explicit acknowledgment when questions cannot be answered.
- **Sample**: dataset instance sample.

F METADATA EVALUATION OF DATASETRESEARCH-PRO

We present the metadata results of six discovery approaches on `DatasetResearch-Pro` through a radar chart, as shown in Figure 7. The chart clearly demonstrates significant performance differences between the two major method categories, particularly highlighting the comprehensive advantages of the Deep Research method in the Search Group across all dimensions, with especially excellent performance in Task and Output Alignment. The Deep Research approach brings comprehensive benefits, particularly in task and output aspects. Notably, generative methods perform most prominently in the Output and Intro dimensions, reflecting the inherent advantages of synthetic data in output quality and instruction-following accuracy for tasks. Furthermore, the comparison between OpenAI o3 w/ ref and w/o ref validates the positive impact of reference samples on improving data generation quality.

G PROMPTS

G.1 PROMPTS FOR DATASET CURATION

This prompt instructs OpenAI o3 to generate a comprehensive dataset demand description by transforming extracted metadata into generic, discoverable descriptions that omit specific identifiers while preserving all essential characteristics needed for effective dataset retrieval.

Prompt for demand description generation

Imagine you are performing a dataset search task, and the ultimate ground truth (the most ideal dataset found) is the dataset I provided to you. Now, based on your extraction of this dataset's metadata in the previous conversation, generate a prompt for the dataset search task (this prompt will be passed to a powerful deep research agent to complete the task). This prompt needs to include all the metadata from the previous conversation, except for the example.

Previous metadata extraction: {metadata_json}

Please generate a search prompt that a research agent could use to find this exact type of dataset. The prompt should be comprehensive and include all the key characteristics that would help identify this specific dataset type.

IMPORTANT REQUIREMENTS:
1. DO NOT mention the original dataset name or any specific dataset identifiers
2. Include all metadata information EXCEPT the example field
3. For dataset size/samples count, use approximate ranges rather than exact numbers (e.g., "around 10K samples" instead of exact counts)
4. Focus on the task type, domain, input/output characteristics, and source information
5. Make the prompt generic enough that it could find similar datasets, not just the specific one

Output the search prompt directly without any additional formatting or explanation.

G.2 PROMPTS FOR DATA RESEARCHER

This prompt instructs AI to search Hugging Face for publicly accessible datasets and return a specified number of dataset IDs in strict JSON format.

Prompt for search agent

{agent_query}

IMPORTANT: You must search for publicly accessible datasets from Hugging Face and return exactly {NUM_SEARCH} suitable dataset IDs.

SEARCH STRATEGY:
- First, search for datasets that closely match the query
- If you cannot find enough datasets, gradually relax the search criteria to find more potential datasets
- Include both popular and less popular datasets that might be relevant
- Consider datasets with similar tasks, domains, or data types

OUTPUT FORMAT REQUIREMENTS:
You MUST output your response in EXACTLY this JSON format - do not include any other text or explanations:

{'search_datasets': ['dataset_id_1', 'dataset_id_2', 'dataset_id_3', 'dataset_id_4', 'dataset_id_5', 'dataset_id_6', 'dataset_id_7', 'dataset_id_8', 'dataset_id_9', 'dataset_id_10']}

- Use ONLY the exact format above
- Replace dataset_id_1, dataset_id_2, etc. with actual Hugging Face dataset IDs
- Ensure you provide exactly {NUM_SEARCH} dataset IDs
- Do not include any text before or after the JSON
- The JSON must be valid and parseable

This fine-tuning data extraction prompt guides AI to analyze dataset structure and establish conversion rules for transforming it into fine-tuning format with input-output fields, requiring a mandatory task-specific instruction template.

Prompt for fine tuning data extraction from huggingface datasets

You are tasked with analyzing a dataset and determining how to convert it into a fine-tuning format for language models. The fine-tuning data should have exactly two fields: "input" and "output".

{readme_section}
{config_section}
{split_section}{sample_section}

Based on the README content (if available) and the sample data, please analyze and provide the following information:

1. **Selected Config**: Which config/subset was selected from the dataset (output "None" if there's only one config)
2. **Selected Split**: Which split was selected (train/test/validation/etc.)
3. **Conversion Rules**: Describe the rules for converting the dataset samples into input/output format for fine-tuning, including:
- Which columns/fields should be combined for the "input"
- Which columns/fields should be used for the "output"
- **MUST include appropriate instruction text** (e.g., "Please answer the following question:", "Classify the sentiment:", "Translate the text:", etc.)
- The exact format and order for combining the fields
- Any preprocessing needed for the data

IMPORTANT REQUIREMENTS:
- The "instruction_template" field MUST contain a clear, specific instruction appropriate for the task
- The instruction should tell the model exactly what to do (e.g., answer questions, classify, translate, summarize, etc.)
- Do NOT use "null" for instruction_template - always provide a meaningful instruction
- The input should clearly guide the model on what output is required

Please provide your response in the following JSON format:

```
{
"selected_config": "config_name or None",
"selected_split": "split_name",
"conversion_rules":
{
"input_components": ["list of field names to include in input (use exact field names from sample)"],
"output_components": ["list of field names to include in output (use exact field names from sample)"],
"instruction_template": "REQUIRED: Clear instruction text telling the model what to do",
"input_format": "detailed description of how to format the input",
"output_format": "detailed description of how to format the output",
"example_conversion":
{
"input": "example input with instruction based on the sample",
"output": "example output based on the sample"
}
}
}
```

Important: Make sure the conversion makes sense for language model fine-tuning and follows common patterns for instruction-following datasets. The instruction_template is MANDATORY and should be task-specific.

This synthesis agent prompt tasks AI with directly synthesizing a specified number of high-quality training samples based on data requirements, aiming to create superior fine-tuning data compared to existing searched datasets.

Prompt for synthesis agent

You are a specialized expert in fine-tuning data synthesis. You have the following dataset search requirement: {agent_query}

Your task is to directly synthesize {num_data} corresponding examples based on this requirement. The goal is to create synthetic data that, when used for fine-tuning a large language model, will achieve better performance than fine-tuning on existing datasets found through the search.

Here is a reference example for guidance: {example_data}

You MUST output exactly {num_data} samples in JSON list format, where each sample contains only "input" and "output" fields, following this exact format:

{ "input": "...", "output": "..." }, { "input": "...", "output": "..." }, ...

Important requirements:
1. Generate EXACTLY {num_data} examples
2. Each example must have only "input" and "output" fields
3. Follow the task type and domain specified in the search requirement
4. Use the reference example to understand the target format and style
5. Ensure diversity across your generated examples
6. Focus on creating high-quality data that will improve model performance through fine-tuning

G.3 Prompts for Metadata Evaluation

This prompt instructs AI to analyze a HuggingFace dataset's README and sample data to extract comprehensive metadata including task type, domain, input/output descriptions, and source information in structured JSON format.

Prompt for dataset metadata generation

You will see a README file introduction from a HuggingFace dataset and one sample from it. You need to output the following content based on these materials. Please output in JSON format.

README content:
{readme_content}{sample_section}

Please analyze the dataset based on both the README and the sample (prioritize the sample when there are conflicts, as the README might be vague while we've ensured all samples in the dataset are similar to the provided one), and output the following metadata in JSON format:

{
"introduction": "A one-sentence introduction of the dataset content, concise and clear, including key information about task, domain, input and output",
"task": "Directly output one of: multiple-choice, question-answering, summarization, text-classification, text-generation, translation",
"domain": "Directly output the domain the dataset content involves, such as: aerospace, finance, linguistics, politics, sociology, biology, etc.",
"input": "Directly output the dataset's input content, including its language, such as: an English news text for translation, a multiple-choice question in French philosophy domain, etc. (Consider both sample and README, don't be limited by single sample's domain, but also don't be too broad like README)",
"output": "Directly output the dataset's output content, including its language, such as: a number 0 or 1, a letter A/B/C/D, translated Italian text, etc.",
"source": "Directly output the dataset's source: real-world, human-generated, machine-generated, etc.",
"example": "Directly extract the sample provided in the prompt and put it here",
"samples_count": {samples_count}
}

Important: Please strictly follow the above JSON format and provide a comprehensive analysis based on both README and sample data.

This metadata evaluation prompt tasks AI with comparing two dataset metadata objects across multiple dimensions and providing numerical similarity scores (0-10) for each dimension along with an overall average score in JSON format.

Prompt for metadata evaluation

I need you to compare two dataset metadata and score their matching degree across the following dimensions.

Dimension descriptions:
- introduction: Dataset introduction and overview
- task: Task type (e.g., text-classification, question-answering, etc.)
- domain: Domain field (e.g., finance, politics, biology, etc.)
- input: Description of input content
- output: Description of output content
- source: Data source (e.g., human-generated, machine-generated, etc.)
- example: Sample data
- samples_count: Number of samples

Original dataset metadata:
{original_metadata}

Search dataset metadata:
{search_metadata}

Please score each dimension on a scale of 0-10 for matching degree, where:
- 10 points: Complete match or highly similar
- 0 points: Complete mismatch or opposite
- Output an integer score. If a dimension is missing or meaningless in one or both metadata, mark it as null

Please output the result strictly in the following JSON format:

{
"introduction": score or null,
"task": score or null,
"domain": score or null,
"input": score or null,
"output": score or null,
"source": score or null,
"example": score or null,
"samples_count": score or null,
"average": average score (excluding null values) or null
}

Note:
1. Only output JSON format, do not include any other text
2. Scores must be numbers between 0-10 or null
3. average is the mean of all non-null scores

H CONFIG FOR SUPERVISED FINE-TUNING

Below is the standard configuration file for supervised fine-tuning using the LlamaFactory framework (Zheng et al., 2024), based on the Llama-3.1-8B model with bfloat16 precision and full-parameter fine-tuning strategy:

Config for supervised fine-tuning

```
bf16: true
cutoff_len: 4096
dataloader_num_workers: 0
dataset: {dataset_id}
ddp_timeout: 180000000
deepspeed: examples/deepspeed/ds_z3_config.json
do_train: true
finetuning_type: full
gradient_accumulation_steps: 2
learning_rate: 1.0e-05
logging_steps: 10
lr_scheduler_type: cosine
max_samples: 1000
model_name_or_path: models/LLama3/Llama-3.1-8B
num_train_epochs: 3.0
output_dir: LLaMA-Factory/results/{task_id}/saves
overwrite_cache: true
overwrite_output_dir: true
per_device_train_batch_size: 1
plot_loss: true
preprocessing_num_workers: 16
report_to: none
resume_from_checkpoint: null
save_only_model: false
save_steps: 1000
stage: sft
template: llama3
trust_remote_code: true
warmup_ratio: 0.1
```

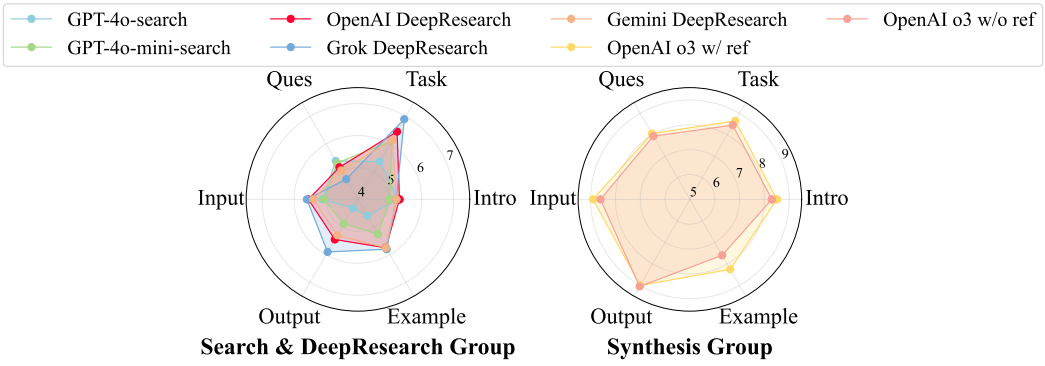

Figure 7: **Radar chart illustrating the average metadata alignment scores for the Search Group, Synthesis Group, and DeepResearch Group on the `DatasetResearch-Pro` subset.** The Generation and DeepResearch groups achieve higher and more balanced scores, indicating superior semantic alignment with the ground-truth dataset requirements.

I Analysis

This section delves into specific cases to provide a qualitative understanding of the performance patterns observed in our experiments. By examining individual examples from Figure 8, Figure 9 and Figure 10, we highlight the distinct behaviors of search, synthesis, and deep research methodologies,

DataResearcher	Method	DTP Score		Meta Score						
		Knowledge(%)↑	*Reasoning*(%)↑	Intro	Task	Ques	Input	Output	Example	Avg ↑
Baseline		10.41	11.84							
Search-based										
GPT-4o-mini-Search	1-shot	6.48	5.73	5.50	6.44	5.29	5.70	4.88	5.33	5.52
	3-shots	6.45	7.89							
	5-shots	10.38	8.39							
	Fine-tune	12.12	17.35							
GPT-4o-Search	1-shot	9.82	7.25	5.76	6.63	5.36	5.77	5.32	5.41	5.71
	3-shots	9.84	8.43							
	5-shots	10.22	8.70							
	Fine-tune	41.89	27.54							
GPT-5	1-shot	20.15	15.84	7.28	8.04	6.99	7.35	6.80	7.10	7.26
	3-shots	20.54	16.33							
	5-shots	22.34	18.81							
	Fine-tune	45.29	34.52							
Synthesis-based										
Deepseek-V3 w/ ref	1-shot	8.97	13.36	8.23	8.46	7.96	8.38	8.67	7.80	8.25
	3-shots	13.32	20.18							
	5-shots	14.58	21.36							
	Fine-tune	28.16	49.15							
Claude-4-Sonnet w/ ref	1-shot	9.86	15.89	8.63	8.51	8.56	8.03	8.83	9.22	8.21
	3-shots	19.93	26.15							
	5-shots	20.28	28.25							
	Fine-tune	35.14	65.16							
GPT-5 w/ ref	1-shot	12.44	24.86	8.88	9.02	8.33	9.30	9.69	8.60	8.97
	3-shots	23.69	35.12							
	5-shots	26.81	39.46							
	Fine-tune	43.67	75.46							
OpenAI o3 w/ ref	1-shot	10.25	17.38	8.63	8.71	8.11	9.01	9.36	8.32	8.69
	3-shots	21.81	32.14							
	5-shots	23.91	28.92							
	Fine-tune	38.98	72.70							
OpenAI o3 w/o ref	1-shot	10.16	12.26	8.58	8.74	8.11	8.80	9.17	8.04	8.57
	3-shots	17.25	25.53							
	5-shots	14.81	19.44							
	Fine-tune	37.94	67.25							

Table 5: **Main Results of DATASETRESEARCH.** We report normalized few-shots and fine-tune scores on Knowledge and Reasoning tasks using DTP (Downstream Task Evaluation), alongside metadata similarity scores.

particularly focusing on the fine-tuning performance. Full main experimental results is shown in Table 5.

I.1 KNOWLEDGE VS. REASONING

Figure 8 contrasts how different methods handle knowledge and reasoning tasks (Ravichander et al., 2019). For reasoning-centric requirements, synthesis can construct detailed data with explicit reasoning paths, consistent with its large fine-tuning advantage. For knowledge tasks, retrieval provides broad factual coverage and yields the clearer few-shot advantage, although search and synthesis are close under fine-tuning.

This reveals a key behavioral pattern: the breadth advantage of search methods allows them to capture long-tail knowledge distributions, making models more robust for diverse, fine-grained factual queries. Conversely, the depth advantage of synthesis methods lies in the ability of powerful LLMs to construct highly structured, logically coherent reasoning data, making it ideal for training models on tasks that depend on generalizable logic rather than pure fact retrieval.

I.2 DEEP RESEARCH METHODOLOGY

As observed in Figure 9, the deep research methodology, with its iterative information gathering and reasoning-guided exploration, retrieves data of significantly higher quality than single-shot search or generation. Under identical query conditions (Dao et al., 2023), the resulting dataset format demonstrates superior alignment with the reference set's characteristics, exhibiting greater



Figure 8: **Case study comparing data construction for Reasoning-intensive versus Knowledge-intensive tasks.** The left example (a legal reasoning task) shows that a synthesis agent generates a structured, high-quality output. The right example (an Arabic cultural classification task) demonstrates that a search agent successfully finds a highly relevant existing dataset with broad factual coverage.

comprehensiveness and analytical depth. This core advantage stems from its multi-round process, which constructs a more refined analytical perspective and uncovers high-quality datasets that are otherwise difficult to find. This enhances not only the breadth of information but, more importantly, the logical coherence of the resulting dataset, leading to clear performance gains in the final fine-tuning evaluation.

I.3 Highly Specialized Tasks

Figure 10 illustrates a highly specialized task for which the evaluated search and synthesis methods find little relevant coverage (Shoham & Rappoport, 2024). The resulting fine-tuning performance is poor, but this case alone does not establish that useful reconstruction is impossible.

Search is limited by indexed sources, while synthesis depends on learned patterns. The RTS, oracle, and human results show that better selection and transformation can recover additional signal, yet the remaining human–reference gap indicates that some target-specific information may not be available from the demand and web alone.

Deep Research Case

Reference Set Sample (Accuracy: 0.33)
dataset_id: vnhsge/biology
Input: Câu 81. Trên một NST thường, khoảng cách giữa hai gen A và B là 34cM. Theo lí thuyết, tần số hoán vị giữa hai gen này là\nA. 33%.\t\nB. 17%.\t\nC. 66%.\t\nD. 34%.
Output: D

Dataset Demand :
Find a dataset for supervised fine-tuning that contains Vietnamese national exam-style questions spanning multiple academic subjects (literature, English, history, geography, civics, STEM) and includes both multiple-choice questions and literary essays. The dataset should feature multi-modal inputs (text and accompanying images) and use an instruction-input-output format where each sample includes a system directive, an input field with exam texts (and potential image references), an output field with a multiple-choice answer (e.g., a letter), and an output_program field providing detailed reasoning. It must support multi-task evaluation (question answering, text generation, reading comprehension, and visual question answering) and be large-scale (from 10K up to 100K samples) so that it benchmarks LLM performance against high-stakes national exam conditions with domain-specific reasoning and long-context handling.

OpenAI o3 Synthesis Dataset Sample (Accuracy: 0.05)
Input: <<SYSTEM>> Bạn là trợ lý luyện thi THPT Quốc Gia. Trả lời theo yêu cầu.\n### Bài đọc hiểu (Ngữ văn):\nĐọc đoạn trích sau và trả lời câu hỏi.\n \"...Những đêm sương mùa đông lạnh buốt, Mị vẫn nghe tiếng sáo gọi bạn, tiếng chó sủa làng xa vang vọng trên nương. Trong hơi men rượu ngô, Mị tưởng như mình đang được đi chơi tết ngày trước...\" (trích Vợ chồng A Phủ – Tô Hoài)\nCâu hỏi: Tác dụng nghệ thuật nổi bật của chi tiết \"tiếng sáo gọi bạn\" trong đoạn trích là gì?\nA. Khắc họa bối cảnh thiên nhiên vùng cao.\nB. Thể hiện khát vọng sống, khát khao tự do của nhân vật Mị.\nC. Tái hiện không khí lễ hội tết cổ truyền.\nD. Tô đậm hoàn cảnh nghèo khổ, lạc hậu của người Mông.\n”
Output: Đáp án: B\n\nLời giải chi tiết: \"Tiếng sáo gọi bạn\" là một motif âm thanh quen thuộc trong văn học Tây Bắc, gợi nhịp sống rộn ràng của tuổi trẻ và tình yêu. Chi tiết ấy đột ngột vang lên giữa đêm đông giá lạnh, đánh thức sâu thẳm trong Mị khát vọng sống mãnh liệt bị kìm nén bấy lâu. Do đó, nó không chỉ miêu tả không khí lễ hội hay bối cảnh, mà chủ yếu là nghệ thuật thức tỉnh tâm hồn nhân vật, thể hiện khát vọng sống, khát khao tự do.

OpenAI gpt-4o Search Dataset Sample (Accuracy: 0.005)
Huggingface dataset_id:
testli/English_exam31.1
Input: Please read the following instruction and its associated difficulty level, then produce the expected output accordingly. Follow the guidelines carefully.\nGenerate 9 multiple-choice questions for the Vietnamese National High School Graduation Examination in English, specifically for the reading comprehension advanced (requires deeper analysis and inference from text) section. Each question should:\n - Include four answer options.\n - Clearly identify the correct option, ensuring it is one of the provided options (i.e., it must be 0, 1, 2, or 3).\n - Ensure that the exam_section value matches 6.\n Additional guidelines:\n 1. Ensure questions and passages are unique and challenging.\n 2. Use grammatically correct English and avoid duplication of passage meanings, topics, or questions.\n 3. Exclude any harmful, inappropriate, or culturally insensitive content.\n4. Ensure the Flesch-Kincaid Grade level is 11\n11
Output: ‘text_section_body’: \"My favorite Vietnamese festival is Tết Nguyên Đán, often called Tet, the Lunar New Year. It’s a time for family reunions, enjoying delicious feasts, and celebrating new beginnings. Preparations begin weeks in advance; families clean their houses meticulously symbolizing washing away the past year‘s bad luck. Red envelopes containing lucky money are prepared for children and the elders exchange greetings and warm wishes. The family gathers for elaborate meals, and traditional lion and dragon dancers parade through the streets carrying joyous celebration through its many activities. This multi-day celebration fills me with warmth and hope every year.\“, \n’question‘: [’question_text_0: \\‘What is the most vital essence of Tet ?\\’,\\noptions_of_question_text_0: [\\‘The festival takes one day to celebrate.\\’, \\‘Homes are lightly cleaned up to represent the season.\\’, \\‘The festival only involves festive meals.\\’, \"It\\’s primarily characterized by gathering family, enjoying feast,and welcoming new beginnings\"],\\ncorrect_option_of_question_text_0: 0,\\n', \"question_text_1: 'What reflects deep aspects of celebrating the Tet festival preparing for its activities ?',\\noptions_of_question_text_1: ['Children enjoy no such celebrations.', 'Preparing takes days but the traditions are few.', 'The money given is not particularly special…

OpenAI DeepResearch Search Dataset Sample (Accuracy: 0.19)
Huggingface dataset_id:
SEACrowd/m3exam
Input: I. ĐỌC HIỂU (3,0 điểm)\nĐọc đoạn trích:\n“chính mẹ đẻ anh hùng và truyền thuyết\ntù túp lều lợp lá lợp tranh\ncất cuống nhau bằng lưỡi liềm\nbàn chân thô quanh năm bùn lấm\nchưa một lần ướm qua sử sách\ntập con bước vịn vào ca dao tục ngữ\ndù uống nước đâu lòng vẫn nhớ nguồn\nthương từ cái kiến con ong\ntìm ruột bầm gan thù bọn ác\ndân tộc tôi khi đứng dậy làm người\nlà đứng theo dáng mẹ\n\"đón gánh tre chín dạn hai vai\" (1)\nmùa hạ gió Lào quăng quật\nmùa đông sắt se gió bấc\ndân tộc tôi khi đứng dậy làm người\nmồ hôi vã một trời sao trên đất\ntrời sao lăn hóa thành muôn mạch nước\nchảy âm thầm chảy dọc thời gian “\n\n(1) Câu thơ của Nguyễn Du. Trích Những người đi tới biển, Thanh Thảo, NXB Quân đội Nhân dân, 2004, tr. 53-54)\nThực hiện các yêu cầu sau:\nCâu 1. Xác định thể thơ của đoạn trích.”
Output: Thể thơ của đoạn trích: Thể thơ tự do

Figure 9: In this case, deep research agent demonstrates superior performance over DataResearcher with search and synthesis agents.

Failed on Corner Case

Reference Dataset Sample (Accuracy: 0.917)
Huggingface dataset_id:
ofir408/MedConceptsQA
Input: Question: What is the description of the medical code 89.43 in ICD9PROC?\nA. Enteral infusion of concentrated nutritional substances\nB. Cardiovascular stress test using bicycle ergometer\nC. Measurement of systemic arterial blood gases\nD. Replacement and removal of therapeutic appliances, option1: Enteral infusion of concentrated nutritional substances, option2: Cardiovascular stress test using bicycle ergometer, option3: Measurement of systemic arterial blood gases, option4: Replacement and removal of therapeutic appliances
Output: Cardiovascular stress test using bicycle ergometer

Dataset Demand :
Search for a dataset designed for few-shot learning evaluation that uses an instruction fine-tuning format with system-input-output triplets. The dataset should consist of detailed multi-choice questions with contextual prompts and enumerated answer options (e.g., A, B, C, D) along with a system directive such as \"Select the best option.\" The expected output must be a short, full-sentence answer that exactly selects the correct option based on domain-specific content, including professional and academic tasks like medical coding (e.g., ICD9PROC). The collection should encompass 1K-10K examples focused on assessing factual accuracy and generalization in low-resource settings.

OpenAI o3 Synthesis Dataset Sample (Accuracy: 0.0)
Input: System: You are a certified medical coding assistant. Select the best option.\n\nContext: A 58-year-old patient underwent a laparoscopic cholecystectomy with intraoperative cholangiography for symptomatic gallstones.\n\nQuestion: Which ICD-9-CM procedure code most accurately describes the primary surgical procedure performed?\nOptions:\nA. 51.22 – Cholecystectomy, not otherwise specified\nB. 51.23 – Laparoscopic cholecystectomy\nC. 51.24 – Open partial cholecystectomy\nD. 51.19 – Other excision of gallbladder\n\nTask: Select the best option.
Output: The correct option is B.

OpenAI gpt-4o Search Dataset Sample (Accuracy: 0.067)
Huggingface dataset_id:
shuyuej/instruction_tuning_data
Input: You are a medical expert. Please read the following instruction and patient case carefully, then select the best answer among the provided options and explain your reasoning in detail.\nYou're a doctor, kindly address the medical queries according to the patient's account.\nAnalyze the question and answer with the best option.\nA 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?\n A. Ampicillin\n B. Ceftriaxone\n C. Doxycycline\n D. Nitrofurantoin
Output: The correct answer is (D).\n\nExplanation: The patient's presentation is indicative of urinary tract infection (UTI), which is common in pregnant women due to urinary stasis that occurs from progesterone-induced ureteral relaxation and compression of the bladder by the growing uterus. UTI in pregnancy is considered complicated by the inherently increased risk of pyelonephritis in this population and is treated with 7 days of antibiotics regardless of symptoms. Nitrofurantoin, as in option D, is first line for UTI in pregnancy due to its safety profile. It's important to note, though, that it's not used in late pregnancy (i.e., at or after 38 weeks of gestation) or in patients with a creatinine clearance of <40-60 mL/min due to hemolytic anemia risk in the neonate and lack of drug concentration in renal tissue, respectively.\n\nAmpicillin (option A) and Ceftriaxone (option B) are not typically first-line options for uncomplicated cystitis. Doxycycline (option C) should be avoided in pregnancy due to the risk of teratogenicity.

Figure 10: Performance comparison in a challenging corner case. All evaluated methods show degraded performance in this scenario.