DeepDiver-DR: Empowering LLMs for Deep Research via Specialized Multi-Agent Collaboration

Anonymous ACL submission

Abstract

Deep Research (DR) systems iteratively plan, retrieve evidence, and synthesize long-form reports in response to complex, open-ended queries. Existing approaches suffer from context overload, role entanglement, and error propagation over long horizons, while the most capable DR systems remain locked behind massive proprietary models. We present **DeepDiver-DR (DD-DR)**, a multi-agent framework that decomposes the DR process into three specialized roles—planner, information seeker, and writer—coordinated through a persistent, file-based shared workspace. Each role is further equipped with fine-grained, DR-specific tools. To support rigorous evaluation, we introduce **RubricWrite-Bench (RW-Bench)**, a benchmark of 100 rubric-based writing queries with a two-phase assessment of global report quality and point-level rubric fulfillment. Experiments show that DD-DR(30B) establishes a new state of the art among models at the same parameter scale on the DeepResearch Bench, and achieves performance competitive with leading proprietary DR systems on RW-Bench. Ablations further confirm the contribution of the architectural components.

1 Introduction

Large language models (LLMs) (OpenAI, 2025b; Yang et al., 2025; Comanici et al., 2025) have achieved strong performance on tasks ranging from mathematical reasoning (Ahn et al., 2024) to code generation (Jiang et al., 2026), but many real-world problems demand more than short, one-off answers. They require the critical synthesis of factually grounded information into coherent, self-contained deliverables. The Deep Research (DR) paradigm (Google, 2025) addresses this need: given a complex, open-ended query, a DR system iteratively plans sub-questions, acquires evidence from external sources, and synthesizes its findings into a structured, long-form report.

Prior approaches to DR fall short at progressively deeper levels. Static Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) pipelines are confined to pre-indexed corpora and cannot interact with dynamic sources such as search engines or live web pages. Single-agent systems (Russell et al., 1995) address this problem but suffer from role entanglement—planning, retrieval, and writing compete for the same context window—leading to error propagation and state drift. Multi-agent systems (MAS) (Wooldridge, 2009) distribute responsibilities across roles; however, existing MAS-based DR frameworks still exhibit critical weaknesses: coarse role decompositions that misalign with the DR pipeline, communication protocols that forward full conversation histories and trigger context overload, and generic tooling that lacks the precision required for evidence-grounded writing. For example, WebWeaver (Li et al., 2025b)—a popular DR system—employs a dual-agent design where the planner jointly handles evidence acquisition and outline optimization; however, coupling retrieval with planning in a single agent constrains parallel evidence gathering, and the absence of dedicated quality-control tools limits the reliability of the generated report. Moreover, leading DR systems (OpenAI, 2025a; Google, 2025; Anthropic, 2025a) depend on massive proprietary models, motivating frameworks that can empower open-source, smaller-scale LLMs.

We present DeepDiver-DR (DD-DR), a multi-agent framework that addresses these gaps through three design choices. First, we decompose the DR process into three specialized roles—planner, information seeker, and writer—whose responsibilities map directly onto the core stages of planning, evidence acquisition, and report synthesis (Shi et al., 2025b). Second, we replace full context forwarding with a shared-workspace mechanism where agents exchange only compressed task summaries and persist intermediate artifacts as files, keeping

per-agent context short and reducing hallucination risk. Third, we equip each role with DR-specific tools (structured search, document analyzers, and section-level drafting/verification modules), integrated via the Model Context Protocol (MCP) (Anthropic, 2024).

To enable rigorous evaluation, we also introduce RubricWrite-Bench (RW-Bench), a suite of 100 rubric-based writing queries across 10 domains together with a two-phase evaluation protocol. Unlike purely open-ended queries that yield unstable assessments, each query specifies knowledge-intensive writing points that ground both rubric-level scoring and holistic report assessment (Section 4.2.3). Experiments on the evaluation benchmarks show that DD-DR (30B) establishes a new state of the art (SOTA) among models at the same parameter scale and achieves performance competitive with leading proprietary DR systems.

In summary, this paper makes the following contributions: 1) We propose DD-DR, a multi-agent DR framework featuring role decomposition, file-based shared-workspace communication, and role-specific tools that enable open-source LLMs to tackle long-horizon deep research tasks. 2) We introduce RW-Bench, a rubric-based DR benchmark and evaluation protocol for assessing global report quality and coverage of specified writing points. 3) We conduct comprehensive experiments and ablation studies demonstrating competitive performance and validating the contribution of the architectural components.

2 Preliminaries

2.1 Deep Research (DR) Paradigm

Given an open-ended research query $q \in \mathcal{Q}$, a DR system produces a coherent, source-grounded report $y \in \mathcal{Y}$ by interacting with external information sources and tools. Rather than generating y in a single pass, the system operates through an iterative loop of (1) planning or refining sub-questions, (2) acquiring evidence from heterogeneous sources, and (3) synthesizing findings into progressively improved reports.

2.2 Problem Formulation: Multi-Agent DR

We formulate Multi-Agent DR as a long-horizon collaborative process in which N specialized agents, with policies $\{\pi_{\theta_i}\}_{i=1}^{N}$, jointly produce a report for a query $q \in \mathcal{Q}$. Each agent i maintains a private history $h_{t-1}^{(i)}$ and exchanges messages with other agents. At step t, agent i selects an action from a space $\mathcal{A}$ comprising **tool actions** and **communication actions**:

$$a_t^{(i)} \sim \pi_{\theta_i}\Big(\cdot \mid h_{t-1}^{(i)},\, s_t,\, \mathcal{M}_t^{(i)}\Big) \tag{1}$$

where s_t is the environment state, $\mathcal{M}_t^{(i)} = \{m_t^{(k\to i)}\}_{k\neq i}$ is the aggregated incoming messages, and the private history is updated as $h_t^{(i)} = h_{t-1}^{(i)} \cup \{a_t^{(i)}, o_t^{(i)}\}$ with observation $o_t^{(i)}$. After at most T steps, the system generates a report:

$$y = G\Big(s_T,\, \{h_T^{(i)}\}_{i=1}^{N}\Big)\,, \tag{2}$$

which should be responsive to q, well-structured, and grounded in the evidence accumulated in s_T. The design objective is to maximize report quality under given resource constraints.

3 Methodology

Our framework decomposes the long-horizon DR process into specialized sub-tasks handled by three collaborating agents: a **planner**, one or more **information seekers**, and a **writer** (Figure 1). The planner orchestrates query decomposition and dispatch, information seekers acquire and evaluate evidence in parallel, and the writer synthesizes curated materials into a structured report. All agents follow the ReAct paradigm (Yao et al., 2022) and coordinate through a persistent file-based shared workspace.

3.1 Agent Role Design

Planner. The planner is responsible for global reasoning, workflow management, and quality control. It decomposes the research query into a hierarchical task tree, dynamically dispatches sub-tasks to downstream agents, and evaluates whether the gathered evidence is sufficient before advancing the workflow. The planner maintains the authoritative project state through versioned planning documents stored in the shared workspace.

Information Seeker. Multiple information seekers can be instantiated concurrently to investigate different sub-topics. Their scope is strictly limited to information acquisition: they retrieve web content, perform multi-dimensional analysis on the extracted documents (assessing timeliness, authority, richness, etc.), and condense the raw data into structured summaries deposited in the workspace.

Writer. The writer operates on the curated materials assembled by the planner and seekers. It generates a chapter-level outline, maps the evidence

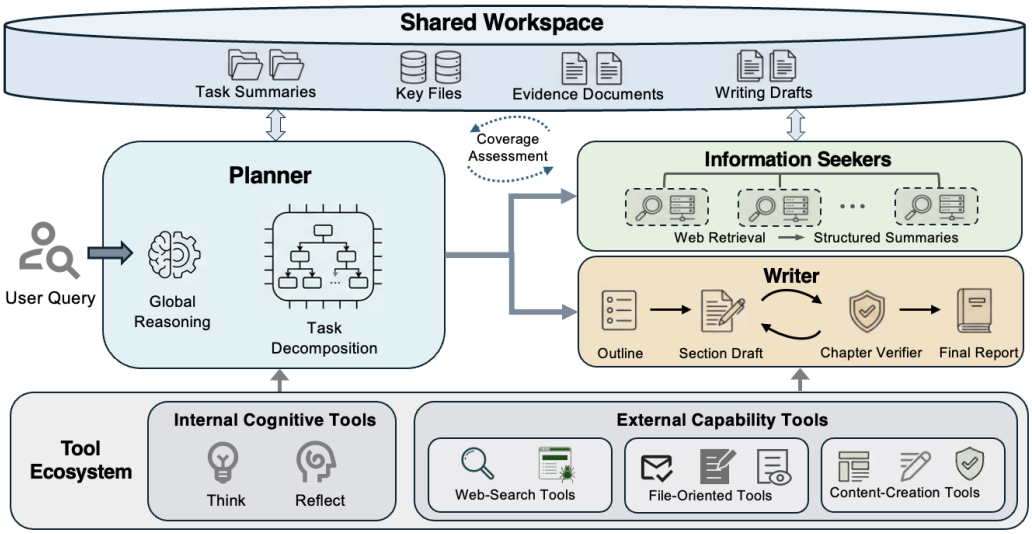

Figure 1: Overview of the DD-DR workflow

to sections, and drafts the report iteratively. After each chapter, a verification step checks logical consistency and factual coverage, triggering a rewrite when the quality falls below the threshold.

3.2 Inter-Agent Communication via Shared Workspace

Rather than forwarding full conversation histories or raw retrieved pages, agents communicate indirectly through a persistent shared workspace $\mathcal{W} = \{f_1, f_2, \ldots, f_k\}$. For agent i with cumulative context $h_{t-1}^{(i)}$ at step t, the message sent to agent j is produced by a communication operator $C_{i \to j}$:

$$m_t^{(i \to j)} = C_{i \to j}\left(h_{t-1}^{(i)}, \mathcal{W}_t\right) = \{\text{task_summary}, \text{key_files}\} \tag{3}$$

The *task summary* is a compressed account of the execution steps, findings, and unresolved issues, which keeps the receiving agent's input concise regardless of how long the sender's reasoning trace was. The *key files* record metadata for every file created or updated during the sub-task, including the file path, content description, and whether the file constitutes a primary deliverable. Agents store all intermediate outputs in the workspace and retrieve specific details on demand through file-oriented tools, rather than carrying them in context.

This design yields three advantages. First, message complexity remains bounded regardless of reasoning history length, ensuring **scalable communication**. Second, all intermediate artifacts are persistently stored and retrievable, providing **durable state management** without maintaining full context windows. Third, independent sub-tasks enable **parallel execution** without state interference.

3.3 Tool Ecosystem

All tools are integrated via the MCP, providing a uniform interface for tool discovery, invocation and result parsing. We organize them into the following two groups:

Internal cognitive tools encapsulate **thinking** and **reflection** as explicit tool calls interleaved with external tool invocations (MiniMax, 2025), helping agents reason over observations before selecting the next action.

External capability tools allow agents to interact with external environments and are organized into three categories:

Web-search tools provide search-engine API access, full-text extraction, format conversion, and multi-dimensional document analysis.

File-oriented tools implement the shared-workspace protocol, supporting file reading, writing, previewing, and LLM-powered document querying in the workspace.

Content-creation tools support sectional drafting through a *search-result classifier*, *section writer*, and *chapter verifier* that maps evidence to outline sections, drafts chapters, and triggers revision when quality is insufficient.

3.4 End-to-End Workflow

The above components are integrated into a four-phase pipeline (Algorithm 1): 1) **Task Planning**: the planner decomposes query q into a hierarchical task tree and writes a planning document to the

Algorithm 1 DD-DR End-to-End Workflow

Require: User query q, workspace $\mathcal{W}$, quality threshold τ
Ensure: Final report y

Phase 1: Task Planning
Analyze q and generate task tree $\mathcal{T} = \{t_1, \ldots, t_n\}$ ▷ planner
Initialize $\mathcal{W}$ with planning document
Phase 2: Information Acquisition
for each $t_i \in \mathcal{T}$ **in parallel do** ▷ seekers
    Retrieve, evaluate, and summarize evidence for t_i, store in $\mathcal{W}$
end for
Phase 3: Iterative Routing
while evidence in $\mathcal{W}$ insufficient **do** ▷ planner
    Refine $\mathcal{T}$ and dispatch additional tasks in parallel ▷ seekers
end while
Phase 4: Document Synthesis
Generate outline O and map evidence to sections ▷ writer
for each chapter c in O **do**
    Draft c and verify quality, redraft if below τ ▷ writer
end for
Merge chapters into y ▷ writer
return y

workspace. 2) **Parallel Acquisition**: information seekers retrieve and summarize evidence for assigned sub-tasks in parallel, returning compressed summaries and `key_files` paths to the planner. 3) **Iterative Routing**: the planner evaluates evidence coverage; if gaps remain, it refines the task tree and dispatches additional seeking rounds. 4) **Document Synthesis**: the writer generates an outline, allocates evidence to sections, drafts each chapter with self-verification, and merges validated chapters into the final report y.

4 Training and Evaluation

4.1 Training Data and Fine-tuning

4.1.1 Query Crafting Pipelines

To construct diverse and challenging training queries, we employ two complementary crafting pipelines. **Internal-Knowledge-Driven Pipeline** provides the model with domain-specific seed keywords and asks it to propose writing topics from parametric knowledge, supporting broad conceptual coverage. **Retrieval-Driven "Inverted" Pipeline** begins with targeted web searches, synthesizes retrieved evidence into draft topics, and iteratively refines them as new evidence is incorporated, yielding externally grounded and contextually faithful queries.

4.1.2 Cold-start Supervised Fine-tuning

A multi-agent DR framework requires its backbone model to follow role-specific instructions, coordinate with other agents, and reliably invoke tools—particularly the file-system operations central to the shared-workspace protocol. To equip the backbone with these capabilities, we construct training queries from multiple domains using the pipelines described in Section 4.1.1, supplemented by in-house queries from additional vertical domains. For each query, we execute the full DD-DR pipeline with Claude-series models (Anthropic, 2025b) as the policy backbones, collecting the resulting planner–seeker–writer trajectories. After careful filtering and post-processing, we curate *DD-DR-2.3k*, a cold-start dataset of 2.3k multi-agent trajectories whose action–observation sequences teach the model to follow role-specific protocols and invoke MCP tools reliably.

4.2 Evaluation Benchmark

4.2.1 Limitations of Existing Benchmarks

Existing DR benchmarks, such as the widely used DeepResearch Bench (Du et al., 2025) and the recent ResearchRubrics benchmark (Sharma et al., 2025), have catalyzed progress but remain two limitations. First, many queries are excessively open-ended; queries such as "Analyze the development of 'God' in various annotations of Laozi" provide insufficient scoping boundaries, leading to divergent reasoning paths and unstable evaluation. Second, evaluator-side specifications are often not fully exposed in the user query. DeepResearch Bench scores open-ended queries against reference reports, while ResearchRubrics provides 101 queries and 2,593 expert-written criteria but only 13 prompts include explicit rubric-like requirements. Thus, most queries remain open-ended while being scored against pre-specified criteria,

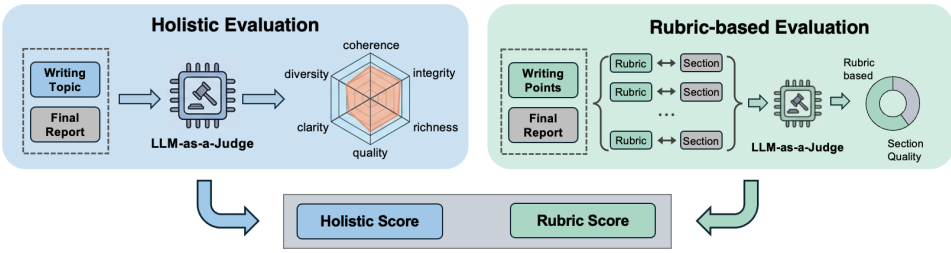

Figure 2: Evaluation framework for the generated final report.

leaving an instruction–evaluation mismatch.

To address these shortcomings, RW-Bench explicitly provides knowledge-intensive writing points that guide report construction and directly ground rubric-based assessment, reducing both open-ended generation variance and instruction–evaluation mismatch.

4.2.2 Benchmark Construction and Quality Assurance

The RW-Bench comprises 100 high-quality writing queries spanning 10 domains. Each query consists of a specific writing topic together with an associated set of writing points that serve as the basis for rubric-based evaluation. The writing points are strictly knowledge-driven, requiring models to retrieve and synthesize external evidence rather than relying on parametric recall. All 100 queries are constructed through rigorous domain selection, expert annotation, and anti-hacking filtering (detailed in Appendix A).

4.2.3 Evaluation Framework

We adopt an LLM-as-a-Judge approach (Zheng et al., 2023) and propose a two-phase evaluation framework that combines holistic and rubric-based evaluation (Figure 2), capturing both global report quality and fine-grained point coverage. For robustness, both phases use three independent LLM judges (`gpt-5.2`, `gemini-3-pro`, and `gemini-3-flash`) and report their averaged scores.

Holistic evaluation. Each judge evaluates the full report across six dimensions: logical coherence, structural completeness, information richness, argumentative quality, language clarity, and content diversity. Each dimension is scored from 1 to 10; their mean yields the holistic score.

Rubric-based evaluation. For each writing point, we first identify the most relevant report sections using BM25 (Robertson et al., 1994); the judge then scores each selected section on a scale of 1 to 10, assessing two aspects: the degree to which the section fulfills the requirements specified by the writing point, and the overall writing quality of the section in terms of fluency, logical coherence, and content diversity.

5 Experiments

5.1 Setup

Benchmark Datasets. We evaluate DD-DR on the following two benchmarks.

DeepResearch Bench (DR Bench) comprises 100 research queries crafted by domain experts across 22 fields. Following previous work (Shao et al., 2025; Anthropic, 2025a), we only report the RACE score on this bench since the FACT score is undermined by unreliable source verification and citation-counting behavior that fails to assess citation relevance or penalize unsupported claims. RACE employs the LLM-as-a-Judge (Gemini-2.5-Pro) to score each generated report against a reference report on four dimensions: Comprehensiveness (Comp.), Insight/Depth (Depth), Instruction-Following (Inst.), and Readability (Read.).

RW-Bench is our proposed benchmark, evaluated with the two-phase rubric-based and holistic evaluation described in Section 4.2.3.

Compared Systems. We compare DD-DR against four leading proprietary DR systems—openai deepresearch (OpenAI, 2025a), gemini-2.5-pro deepresearch (Google, 2025), claude-research (Anthropic, 2025a), and doubao deepresearch—as well as two current SOTA open-source systems at the 30B parameter and 8B parameter scale: WebWeaver (Qwen3-30B-A3B-Instruct-2507) (Li et al., 2025b) and DR Tulu-8B (Shao et al., 2025), for which we report both the SFT and RL variants.

Implementation Details. DD-DR is instantiated with three backbone models: Qwen3-4B-Instruct-2507, Qwen3-8B, and Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025); denoted DD-DR(4B), DD-

DR(8B), and DD-DR(30B), respectively. Starting from the base checkpoints, each variant is fine-tuned on the *DD-DR-2.3k* dataset described in Section 4.1.2, sharing the same trajectory corpus and training recipe and differing only in the backbone capacity.

5.2 Main Results

Results on DR Bench. Table 1 presents the RACE scores on DR Bench. DD-DR(30B) achieves an overall score of 47.77, establishing a new SOTA among open-source models at the same parameter scale and surpassing WebWeaver(30B) by 1.0. It also outperforms openai deepresearch (46.45), claude-research (45.00), and doubao deepresearch (44.34), trailing only gemini-2.5-pro deepresearch (49.71). Beyond the overall score, the per-dimension score reveals a distinctive strength profile. DD-DR(30B) achieves the second-highest Comp. and Depth scores, indicating strong evidence gathering by the dedicated seekers. However, Read. remains its weakest dimension, trailing WebWeaver; we attribute this to the high proportion of structured tool-calling tokens in the SFT training data (see Appendix B for a detailed per-dimension discussion). At the 8B scale, DD-DR(8B) scores 43.63 overall, surpassing both DR Tulu-8B(SFT) (39.00) and DR Tulu-8B(RL) (43.40), especially on Comp. and Depth, again confirming the effectiveness of the dedicated seeker. DD-DR(4B) scores 41.63, demonstrating that the framework can produce reasonable-quality reports even with compact backbones.

Results on RW-Bench. The queries in RW-Bench are accompanied by numerous knowledge-intensive writing points, posing challenges that exceed the capability of the smaller-scale 4B backbone; we therefore report results only for DD-DR(30B) and DD-DR(8B).

Holistic evaluation. Table 2 reports the holistic evaluation scores. DD-DR(30B) achieves an average score of 7.49, closely matching the best proprietary baseline—openai deepresearch (7.54)—and surpassing gemini-2.5-pro deepresearch (7.33) and doubao deepresearch (7.27). Dimension-wise, DD-DR(30B) ranks first on Structural Int. (7.90) and Info. Richness (7.85), confirming that the sectional writing strategy and seeker-driven retrieval pipeline produce well-organized, evidence-rich reports. The main gap lies in Lang. Clarity, consistent with the readability pattern observed on DR Bench (Appendix B).

Rubric-based evaluation. Table 3 presents the rubric-based evaluation results. DD-DR(30B) achieves an average score of 7.02, ranking second only to openai deepresearch (7.12). On the rubric-based score, which measures point-level fulfillment of writing points, DD-DR(30B) attains 7.28, closely trailing openai deepresearch (7.34) while exceeding all other baselines. The section-quality score of DD-DR(30B) reaches 6.75, comparable to doubao deepresearch (6.74) yet slightly behind the other proprietary baselines. DD-DR(8B) obtains an average of 6.48, still competitive with proprietary systems despite its substantially smaller model scale.

Factuality evaluation. To assess factual reliability, we conduct a claim-level factuality evaluation following VeriScore (Song et al., 2024) and SAFE (Wei et al., 2024). Specifically, we randomly sample 30 verifiable claims from each report, generate three search queries per claim to retrieve supporting evidence, and prompt an LLM judge (Gemini-3-Flash) to classify each claim as *Supported*, *Uncertain*, or *Refuted*. Table 4 shows that DD-DR(30B) and openai deepresearch have comparable grounding, with 79%–81% supported claims and about 10% refuted. DD-DR(8B)'s lower supported rate (76.41%) suggests that smaller models struggle mainly during evidence synthesis rather than retrieval.

5.3 Ablation Study

To isolate the contribution of individual components, we ablate the information seeker agent and two key MCP tools, `document_extract` and `chapter_verifier`, from DD-DR(30B) on DR Bench (Table 5).

Information seeker. The No-information-seeker variant collapses the seeker role into the planner, requiring a single agent to handle both task planning and web retrieval. This change leads to a 0.82 drop in the overall RACE score (47.77→46.95), with the most pronounced degradation on Comp. (−1.42) and Inst. (−1.53). Without a dedicated seeker agent, the planner, now burdened with dual responsibilities, causes the number of search queries to drop sharply; the resulting evidence gathering thus becomes shallower and less systematic, leading to narrower topical coverage and weaker alignment with the original query intent. Interestingly, Read. improves slightly, likely because the reduced volume of retrieved material leaves the writer with less heterogeneous content

	Overall	Comp.	Depth	Inst.	Read.
openai deepresearch	46.45	46.46	43.73	49.39	47.22
gemini-2.5-pro deepresearch	**49.71**	**49.51**	**49.45**	**50.12**	**50.00**
claude-research	45.00	45.34	42.79	47.58	44.66
doubao deepresearch	44.34	44.84	40.56	47.95	44.69
Tongyi-Deep Research (30B)	45.66	44.70	44.20	48.82	44.20
WebWeaver (30B)	46.77	45.15	45.78	49.21	47.34
DR Tulu-8B(SFT)	39.00	36.30	35.30	45.50	39.50
DR Tulu-8B(RL)	43.40	41.70	41.80	48.20	41.30
DD-DR (4B)	41.63	40.74	41.59	45.68	32.60
DD-DR (8B)	43.63	43.61	43.31	46.91	37.52
DD-DR (30B)	47.77	47.61	48.71	49.09	42.16

Table 1: RACE scores on DR Bench. Best results are **bolded**, second-best results are underlined.

	Avg. Score	Logical Coh.	Structural Int.	Info. Richness	Arg. Quality	Lang. Clarity	Content Div.
openai deepresearch	**7.54**	**7.67**	7.68	7.63	**7.29**	**7.46**	**7.54**
gemini-2.5-pro deepresearch	7.33	7.63	7.29	7.23	7.25	7.34	7.28
doubao deepresearch	7.27	7.44	7.49	7.50	7.09	7.25	6.86
DD-DR (30B)	7.49	7.63	**7.90**	**7.85**	7.20	6.84	7.50
DD-DR (8B)	6.92	7.11	7.56	7.58	6.87	5.56	6.84

Table 2: Holistic evaluation scores on RW-Bench. Best results are **bolded**, second-best results are underlined.

to integrate, yielding prose that is more uniform in style and easier to read.

Document Extractor. The `document_extract` tool analyzes each retrieved document for source authority, task relevance, and information richness, and deposits compressed summaries into the shared workspace—forming the foundation of the inter-agent communication protocol described in Section 3.2. Without it (No-document-extract), raw retrieved documents are stored directly in the workspace, re-introducing the context overload problem. This ablation causes a severe overall drop (47.77→44.42), with Depth suffering the largest decline (−4.69) and Comp. and Read. also degrading substantially (−2.79 and −4.10, respectively). The results confirm that structured document extraction is essential not only for evidence curation but also for maintaining the compact, high-signal workspace.

Chapter Verifier. The `chapter_verifier` tool scores each drafted section and triggers revision when the score falls below a quality threshold. Removing this quality gate (No-chapter-verifier) leads to a 3.59-point overall drop (47.77→44.18): all four RACE dimensions decline, with Read. suffering the steepest fall (−4.12). Without iterative verification, first-draft sections are accepted unconditionally, and errors in coverage or coherence that would otherwise be caught and corrected accumulate across sections. The particularly steep drop in Read. suggests that the verifier plays a dual role: beyond content checking, its feedback loop encourages the writer to produce more polished prose.

No-SFT ablation. We initially attempted to deploy the off-the-shelf backbones directly within the DD-DR framework without any fine-tuning. However, the base models lack reliable adherence to the role-specific protocols and fail to produce MCP tool calls. In practice, the No-SFT variant was unable to complete the multi-agent workflow and therefore could not generate valid research reports. Therefore, we exclude this configuration from the ablation table.

6 Related Work

Deep Research Agents. DR agents, both proprietary and open-source, have recently attracted substantial attention. Proprietary systems, including OpenAI Deep Research (OpenAI, 2025a), Gemini Deep Research (Google, 2025), and Claude Research (Anthropic, 2025a), achieve strong per-

	Avg. Score	Rubric-based Score	Section-quality Score
openai deepresearch	**7.12**	**7.34**	6.90
gemini-2.5-pro deepresearch	6.96	6.83	**7.08**
doubao deepresearch	6.95	7.15	6.74
DD-DR(8B)	6.48	6.98	5.97
DD-DR(30B)	7.02	7.28	6.75

Table 3: Rubric-based evaluation scores on RW-Bench. Best results are **bolded**, second-best results are underlined.

	Supported	Refuted	Uncertain
openai deepresearch	81.87%	7.14%	10.99%
DD-DR(8B)	76.41%	12.04%	11.55%
DD-DR(30B)	79.21%	10.56%	10.23%

Table 4: Factuality evaluation results on RW-Bench.

	Overall	Comp.	Depth	Inst.	Read.
DD-DR (30B)	47.77	47.61	48.71	49.09	42.16
No-information-seeker	46.95	46.19	47.79	47.56	44.63
No-chapter-verifier	44.18	44.10	44.66	46.06	38.04
No-document-extract	44.42	44.82	44.02	46.93	38.06

Table 5: Ablation study on DR Bench. Each row removes one component from the full DD-DR(30B) system.

formance on complex information-seeking and report-generation tasks, but their closed-source nature limits reproducibility and academic adoption (Zhang et al., 2025). Open-source DR agents mainly fall into two categories. *Workflow-based systems* orchestrate general-purpose LLMs through predefined pipelines. GPT Researcher (Researcher, 2025) and LangChain Open Deep Research (LangChain, 2025), for instance, adopt plan-and-execute workflows with self-reflective knowledge-gap identification, but their static pipelines limit adaptability. *End-to-end trained agents* instead internalize research capabilities within the model. Tongyi DeepResearch (Team et al., 2025), Step-DeepResearch (Hu et al., 2025), and DeepDiver (Shi et al., 2025a) improve training for agentic research, yet their single-agent designs entangle planning, retrieval, and writing within one context window. Recent multi-agent systems partially address this issue: DeepResearcher (Zheng et al., 2025) trains a controller–browser dual-agent system, and WebWeaver (Li et al., 2025b) interleaves outline planning with evidence acquisition. However, role entanglement and raw conversational history forwarding remain common, causing context overload over long horizons. DD-DR addresses these limitations through a three-role decomposition, a file-based shared workspace, and role-specific tools.

Deep Research Evaluation Benchmarks. Recent DR benchmarks cover diverse task settings. DeepConsult (Consult, 2025) targets business and consulting tasks. DeepResearch Bench (Du et al., 2025) comprises 100 PhD-level tasks across 22 domains and evaluates report quality and citation accuracy. ResearchRubrics (Sharma et al., 2025) introduces expert-written rubric criteria for fine-grained DR evaluation, but most of its queries remain open-ended and are evaluated against rubrics that are not explicitly presented as writing requirements. ReportBench (Li et al., 2025a) derives questions from expert survey papers and assesses citation accuracy and factual consistency. DeepResearchGym (Coelho et al., 2025) tests 100 real-world complex queries with LLM-based judgments of report quality, while LiveResearchBench (Wang et al., 2025) focuses on dynamic user-centric tasks. Despite their contributions, existing DR benchmarks often rely on overly open-ended queries, leading to divergent reasoning paths and unstable evaluation. RW-Bench complements them by pairing each query with explicit knowledge-intensive writing points and a two-phase evaluation of holistic report quality and point-level fulfillment.

7 Conclusion

We presented DD-DR, a multi-agent deep research framework, together with RW-Bench, a rubric-based knowledge-intensive writing benchmark. Experiments show that DD-DR(30B) achieves strong performance on both DR Bench and RW-Bench. Ablation studies further verify the contributions of different core components. We hope DD-DR and RW-Bench can serve as useful foundations for future DR research.

Limitations

Despite its strong empirical performance, DD-DR has two limitations. First, DD-DR does not include large-scale reinforcement learning. As a result, the framework may not fully exploit feedback-driven optimization for long-horizon planning, tool use, and evidence-grounded writing. Incorporating RL over multi-agent trajectories is an important direction for future work. Second, our evaluation relies substantially on LLM-as-a-Judge protocols. Although we improve robustness by averaging scores from multiple judge models, LLM judges may still introduce biases, inconsistency, or sensitivity to prompt design. Future work should further combine model-based evaluation with human expert assessment and more objective factuality metrics.

Potential Risks

As a report-generation agent, DD-DR may produce outputs that carry risks common to language-based systems. First, generated reports may contain unsupported claims, or subtle factual errors, especially when retrieved sources are incomplete, unreliable, or misinterpreted during synthesis. Second, the system may reproduce biases, subjective framing, or culturally specific assumptions present in web sources or in the backbone model. Third, fluent and well-structured reports can create an impression of authority even when parts of the content are uncertain, which may mislead users in high-stakes domains such as medicine, law, finance, or public policy. Finally, reports may include inappropriate, sensitive, or harmful content if the user query or retrieved evidence involves unsafe topics. These risks highlight citation checking, human oversight, source transparency, and domain-specific safety constraints when deploying DR agents in real-world settings.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237.

Anthropic. 2024. Introducing the model context protocol. Accessed: 2024-11.

Anthropic. 2025a. Claude takes research to new places. Accessed: 2025-04.

Anthropic. 2025b. Meet claude.

João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and 1 others. 2025. Deepresearchgym: A free, transparent, and reproducible evaluation sandbox for deep research. *arXiv preprint arXiv:2505.19253*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Deep Consult. 2025. Deep consult.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Google. 2025. Gemini deep research — your personal research assistant. Accessed: 2025-03.

Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, and 1 others. 2025. Step-deepresearch technical report. *arXiv preprint arXiv:2512.20491*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2026. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35(2):1–72.

LangChain. 2025. Open deep research.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. 2025a. Reportbench: Evaluating deep research agents via academic survey tasks. *arXiv preprint arXiv:2508.15804*.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, and 1 others. 2025b. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*.

MiniMax. 2025. Interleaved thinking unlocks reliable minimax-m2 agentic aapability.

OpenAI. 2025a. Introducing deep research. Accessed: 2025-02.

OpenAI. 2025b. Introducing openai o3 and o4-mini.

GPT Researcher. 2025. Gpt researcher. Accessed: 2025-02.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Stuart Russell, Peter Norvig, and Artificial Intelligence. 1995. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25(27):79–80.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, and 1 others. 2025. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, and 1 others. 2025. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*.

Wenxuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Hanting Chen, Xiaozhe Ren, Yasheng Wang, Lu Hou, and Lifeng Shang. 2025a. Deepdiver: Adaptive web-search intensity scaling via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, and 7 others. 2025b. Deep research: A systematic survey. *Preprint*, arXiv:2512.02038.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouthi, Caiming Xiong, and Shafiq Joty. 2025. Liveresearchbench: A live benchmark for user-centric deep research in the wild. *arXiv preprint arXiv:2510.14240*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431.

A Benchmark Construction Details

Domain Selection and Filtering. We define a seed set of domains in which 70% overlap with typical LLM training distributions and 30% are out-of-distribution (OOD). For each domain we generate at least five candidate queries using the crafting pipelines described in Section 4.1.1, then apply rigorous manual filtering to retain at most three per domain.

Expert Annotation. A panel of eight expert annotators, all holding a master's degree or higher, evaluates the candidate queries along three dimensions: clarity and fluency, knowledge intensiveness, and the accessibility of the required knowledge from publicly available web sources.

Anti-Hacking Mechanisms. To mitigate data contamination, we employ two countermeasures. For queries derived from documents, we ground them in private sources whose content cannot be retrieved directly but can be reconstructed by synthesizing multiple public web pages. For queries derived from web-based information, we require that each query integrate evidence from at least three distinct websites, preventing single-source retrieval shortcuts.

B Per-Dimension Analysis

Search depth and breadth. On DR Bench, DD-DR(30B) achieves substantial gains over WebWeaver(30B) on Comp. (47.61 vs. 45.15) and Depth (48.71 vs. 45.78); we attribute this improvement to the dedicated information seeker agent, which conducts multi-round, targeted retrieval under the guidance of the planner's structured outline. Unlike single-agent pipelines that interleave search and writing, this separation allows the seeker to pursue deeper and broader evidence gathering without being constrained by the generation context. The same pattern holds at the 8B scale: DD-DR(8B) outperforms DR Tulu-8B(RL) on both Comp. (43.61 vs. 41.70) and Depth (43.31 vs. 41.80), despite using the same backbone model, confirming that the seeker-driven architecture rather than model capacity is the primary driver of evidence coverage. A similar trend emerges on RW-Bench, where DD-DR(30B) obtains the highest scores on Info. Richness (7.85) and Content Div. (7.50), further confirming that the seeker-driven retrieval pipeline enriches the final report with diverse and substantive evidence.

Structural quality. DD-DR(30B) scores competitively on Inst. in DR Bench (49.09) and achieves the best Structural Int. on RW-Bench (7.90). These results reflect a deliberate design choice: rather than asking the writer to produce the entire report in a single run, we equip it with a `section_writer` tool that composes the report one section at a time according to the outline. This sectional writing strategy sidesteps the well-known degradation of coherence and instruction adherence in very long generation contexts, and ensures that each section receives focused attention for logical organization and topical alignment.

Readability gap. Despite strong performance on content and structure dimensions, DD-DR exhibits noticeably lower Read. scores on DR Bench (42.16 for DD-DR(30B), 37.52 for DD-DR(8B)), and a similar Lang. Clarity gap on RW-Bench (6.84 for DD-DR(30B) vs. 7.25–7.46 for proprietary baselines). We hypothesize that this deficit originates from the SFT training stage: the DD-DR trajectory data contains extensive MCP tool invocations formatted in the Hermes JSON schema, and the high proportion of structured, non-natural-language tokens in the training mix may erode the base model's prose fluency. Mitigating this trade-off—for instance through a post-training stage that specifically targets language quality, or by reducing the ratio of tool-calling tokens in the SFT mixture—is an important direction for future work.

C RW-Bench Writing Queries

RW-Bench contains 100 queries in a mixture of Chinese and English. We present examples for each language below.

D Report Writing Case Study

The following case study demonstrates the effectiveness of DD-DR(30B) on long-form report writing for a representative query from RW-Bench. The generated report exhibits strong structural clarity and closely follows the requirements specified in the writing prompt. A key strength of the output is its extensive use of factual evidence and descriptive details retrieved from external sources, enabling a level of granularity and accuracy that cannot be achieved by relying solely on the model's parametric knowledge. This example highlights DD-DR's strong capabilities in information seeking and evidence-grounded content generation.

Example of Writing Query (Chinese)

Writing Topic:

请撰写一篇题为“甜与咸的对话：豆腐脑/豆腐花南北风味分化的历史追踪与社会认同”的综合研究报告，内容需系统覆盖以下要点：

Writing Points:

1. 历史渊源与名称分布考证：追溯豆腐脑（咸系）与豆腐花（甜系）的起源、命名沿革及南北传播路线，并提供关键时间节点与文献证据。
2. 工艺流程与配料差异系统对比：比较甜、咸两大派系在凝固剂、调味料、配料及上桌方式上的技术要点，以图表或流程示意呈现。
3. 味型分化与地域环境、经济因素关联分析：联系气候、主粮结构、糖料与酱料供应及消费经济水平，阐释甜咸分化形成与演变的多重动因。
4. 社会身份与文化认同的象征意义：分析不同人口群体如何通过甜/咸豆腐脑表达地域归属、代际记忆与文化身份，并讨论媒体话语的再生产作用。
5. 营养价值与健康讨论：对比甜、咸版本在能量、糖分、钠含量及添加配料上的营养差异，并梳理公众健康认知与科研结论。

Example of Writing Query (English)

Writing Topic:

Please draft a detailed project proposal for an open-source, modular toolkit that supports the end-to-end creation, training, and assessment of Large-Language-Model-based autonomous agents across multiple domains. Your proposal should cover no more than the following seven key elements:

Writing Points:

1. Core architecture blueprint: specify how the toolkit will let developers configure profiling, memory (structures & operations), planning (with/without feedback), and action modules, ensuring loose coupling and easy extension.
2. Capability-acquisition workflows: outline integrated pipelines for (a) parameter fine-tuning with human, LLM-generated, or real-world data sets and (b) prompt/mechanism engineering without fine-tuning.
3. External-tool plug-in ecosystem: describe interfaces for connecting APIs, databases, knowledge bases, and specialist models, plus a cataloguing system that prevents hallucination-driven misuse.
4. Benchmarking suite design: define tasks drawn from social science, natural science, and engineering; list quantitative metrics (effectiveness, human similarity, efficiency) and standard evaluation protocols (simulation, social interaction, multi-task, software testing).
5. Domain-specific configuration language: propose a concise DSL or YAML/JSON schema for declaring agent roles, goals, environments, and safety constraints.
6. Alignment & safety layer: detail mechanisms for role-appropriate value alignment, hallucination detection/mitigation, and customizable guardrails for sensitive use cases.
7. Community roadmap & governance: set milestones for releases, contribution guidelines, testing infrastructure, and plans for long-term sustainability of the toolkit.

Case Study: Roadmap for Apple-Harvesting Robot Systems

Writing Query:
Please develop a detailed five-year R&D and commercialization roadmap for a next-generation selective apple-harvesting robot system that can operate reliably in high-density orchards. Your roadmap must address the following 7 core elements:
1. Target use-case definition and performance goals: Define orchard scenarios (tree architectures, lighting, weather) and set quantitative KPIs such as pick-rate per hour, fruit-damage threshold, and payback period.
2. Perception & localization subsystem: Specify sensor suite (e.g., RGB-D, LiDAR, multi-spectral) and data-fusion algorithms required to achieve robust fruit detection and 3-D pose estimation under occlusion and variable illumination.
3. Manipulation architecture: Propose manipulator kinematics (single, dual, or multi-arm), required DOF, actuator choices, and cycle-time targets, including a modular end-effector concept that accommodates variable fruit size while minimizing bruising.
4. Autonomous navigation & mobility: Outline the mobile platform, positioning strategy (GPS, vision-based row following, sensor fusion), obstacle avoidance, and orchard fleet-management considerations.
5. Energy strategy and sustainability: Compare on-board batteries versus hybrid solar-assisted power, estimate daily energy demand, and design a charging or swap plan that supports continuous operation.
6. AI-driven planning & control: Describe learning frameworks (e.g., deep reinforcement learning for grasp sequencing) and edge-compute requirements for real-time perception, path planning, and adaptive force control.
7. Commercialization, cost model & risk mitigation: Present a staged budget, BOM cost targets, field-testing milestones, regulatory and safety compliance plan, and mitigation tactics for key technical and market risks (e.g., adverse weather, component costs, grower adoption).

Generated Report (DD-DR(30B))

 # 2026–2030 Apple-Harvesting Robot Roadmap: R&D to Commercialization

**Overview and Scope**
This document defines high-density orchard harvesting targets and builds a complete robotic system architecture for selective apple picking. It establishes row geometry and trellis spacing constraints for systems like...

**Chapter 1. Target Use-case Definition and Performance Goals**

**Row Geometry and Planning Envelopes (High-Density Orchards)**
High-density orchard systems, including the Tall Spindle and Slender Axis systems, are predominant due to their superior profitability and early yield. These systems require precise row spacing and planting densities to accommodate mechanization while ensuring optimal tree growth and light interception. For instance, an ideal in-row spacing for high-density apple orchards such as Tall Spindle or Slender Axis ranges from 3 to 4 feet (approximately 0.9 to 1.2 meters), depending on orchard density requirements...

**Brix Targets (12–18) and Ripeness-Selection Rigor**

885

Apple quality is often measured through soluble solids content (Brix), strongly correlating with consumer preference for sweetness. Defining ripeness selection thresholds exclusively based on visual cues is insufficient for rigorous automation and prevents unnecessary late-picking forces...

Chapter 2. Perception & localization subsystem

Sensors and modalities: RGB-D, LiDAR, IR/UWB; occlusion-aware design
...Recent RGB-D sensing devices, such as the Intel Realsense D435i, provide integrated structured-light or depth-from-stereo sensors capable of producing synchronized color images alongside dense depth maps within a 0.3m to 3m range, with $< 2\%$ depth measurement error at 2m absolute distance...

SGNSS-denied orchard positioning
...Point cloud registration sequences maximize canopy overlap across traverses, achieving mean averagePose errors of approximately 1.173m with standard deviation of 0.498m versus known ground-truth LiDAR trails in shade-dense environments....

Chapter 3. Manipulation architecture and end-effectors

Single vs dual/multi-arm trade-offs (DOF, payload, swivel base)
Designing the manipulation subsystem for high-density selective apple harvesting robots in the EU demands careful consideration of the trade-offs between single-arm and multi-arm configurations, as well as...

Modular, compliant grippers for apple size variability with low bruising
For fruits featuring substantial size variability and petal coatings susceptible to bruising, the advancement of modular, compliant grippers has become an essential focus. The major challenge lies in balancing adaptability with...

...