

WEBWEAVER: STRUCTURING WEB-SCALE EVIDENCE WITH DYNAMIC OUTLINES FOR OPEN-ENDED DEEP RESEARCH

006 **Anonymous authors**

007 Paper under double-blind review

ABSTRACT

013 This paper tackles **open-ended deep research (OEDR)**, a complex challenge
 014 where AI agents must synthesize vast web-scale information into insightful re-
 015 ports. Current approaches are plagued by dual-fold limitations: static research
 016 pipelines that decouple planning from evidence acquisition and monolithic gen-
 017 eration paradigms that include redundant, irrelevant evidence, suffering from
 018 hallucination issues and low citation accuracy. To address these challenges, we
 019 introduce **WebWeaver**, a novel dual-agent framework that emulates the human
 020 research process. The planner operates in a dynamic cycle, iteratively interleaving
 021 evidence acquisition with outline optimization to produce a comprehensive,
 022 citation-grounded outline linking to a memory bank of evidence. The writer then
 023 executes a hierarchical retrieval and writing process, composing the report section
 024 by section. By performing targeted retrieval of only the necessary evidence from
 025 the memory bank via citations for each part, it effectively mitigates long-context
 026 issues and citation hallucinations. Our framework establishes a new state-of-the-art
 027 across major OEDR benchmarks, including DeepResearch Bench, DeepConsult,
 028 and DeepResearchGym. These results validate our human-centric, iterative method-
 029 ology, demonstrating that adaptive planning and focused synthesis are crucial for
 030 producing comprehensive, trusted, and well-structured reports.

1 INTRODUCTION

031 Large Language Models (LLMs) (OpenAI, 2025b; Qwen Team, 2025; Liu et al., 2024; DeepMind,
 032 2025; anthropic, 2025) have demonstrated remarkable capabilities across a wide array of well-
 033 defined tasks, from factual question answering (Wei et al., 2025; Mialon et al., 2023) to document
 034 summarization (Zhang et al., 2025) and code generation (Jiang et al., 2024). Their success, however,
 035 has largely been confined to scenarios with clear instructions and ground-truth answers. The true
 036 frontier for autonomous AI lies in transcending these structured problems to tackle the complex,
 037 open-ended challenges that define human-level knowledge work—a process driven by curiosity,
 038 synthesis, and the discovery of novel insights. We term this challenge open-ended deep research
 039 (OEDR). Unlike tasks with ground-truth answers, OEDR requires an agent to independently navigate
 040 and digest a vast corpus of information, often exceeding 100 web pages and PDFs, to form a detailed
 041 report with accurate citations. This represents a monumental challenge, and as shown in Fig. 1, 2,
 042 most proprietary agents fail dramatically on report quality (RACE) and citation accuracy (FACT) of
 043 DeepResearch Bench (Du et al., 2025), highlighting a critical gap we aim to address.

044 Current attempts to tackle OEDR fall into two main categories: proprietary and open-source solutions.
 045 While several powerful proprietary agents show strong performance (OpenAI, 2025a; Research,
 046 2025b;d;a), their prohibitively expensive APIs and restrictive quotas create significant barriers,
 047 limiting widespread adoption and hindering academic research. Consequently, the focus has shifted
 048 towards open-source alternatives, which predominantly follow two paradigms. As shown in Fig. 3,
 049 the first is a straightforward "search-then-generate" approach (Tao et al., 2025; Roucher et al., 2025),
 050 where the agent gathers all information before directly generating a report. This method results
 051 in low-quality, incoherent outputs because it lacks an outline to guide writing. The second, more
 052 sophisticated approach either initializes a static outline to guide searching (Han et al., 2025; Research,
 053 2025e;c) or searches information for outlining (Shao et al., 2024; Xiong et al., 2025). However,

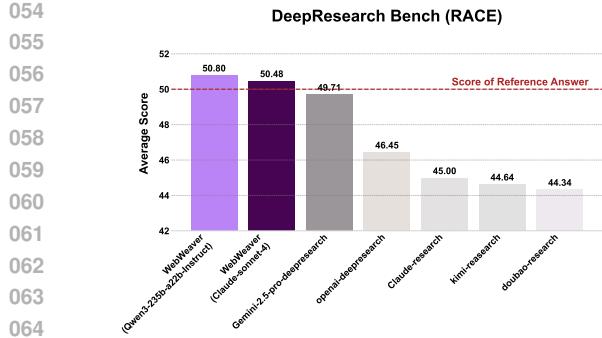


Figure 1: Performance of varying deep research agents on DeepResearch Bench (RACE). The results on DeepResearch Bench are taken from the official leaderboard. Our proposed WebWeaver achieves state-of-the-art performance and even outperforms reference answers.

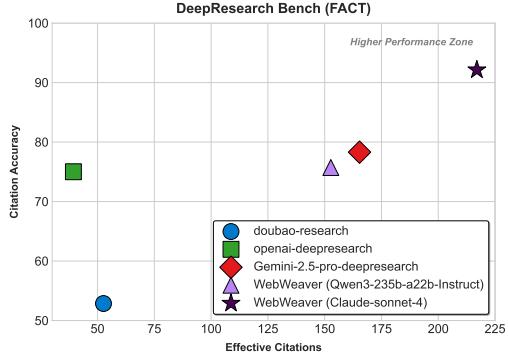


Figure 2: Performance of varying deep research agents on DeepResearch Bench (FACT). Our proposed WebWeaver achieves highest effective citations and citation accuracy.

these strategies are critically flawed: the upfront outline often suffers from the LLM’s internal outdated knowledge. The search-first method also constrains the searching boundary without outline guidance. Furthermore, feeding all searched materials or redundant evidence into the context for final writing is susceptible to the “loss in the middle” issue (Liu et al., 2023) and increased hallucinations, compromising the report’s accuracy and depth (Bai et al., 2024; Wu et al., 2025c).

The key, we believe, lies in abandoning rigid, machine-like pipelines and instead embracing the organic process of human intellect. Our approach is designed to do just that: it teaches the agent to research like a person. A human expert doesn’t decouple and fix their drafting and searching phases; they allow these two phases to co-evolve until converging to a comprehensive outline. We implement this principle through an agentic loop where actions of searching and outline optimization interleave. As the agent explores the web-scale information landscape, its discoveries continuously inform and reshape the outline. Critically, this refined outline then acts as a strategic blueprint, actively guiding subsequent searches to fill identified knowledge gaps and explore underdeveloped sections. This creates a true feedback loop where outlining and discovery co-evolve. Then, when it is time to write, our agent should accept only the relevant context. Just as a human writer would refer to specific notes for a specific chapter, our agent composes each section by focusing only on the source-grounded materials. By doing so, it operates with clarity and precision, crafting a final report that is not just a summary of data but well-structured and insightful pieces of analysis with accurate citations.

To this end, we propose **WebWeaver** by following the human-centric philosophy, a dual-agent framework comprising a *planner* and a *writer*. As shown in Fig. 3, the planner embodies the exploratory research phase, operating in a dynamic, agentic cycle that iteratively interleaves evidence acquisition with outline optimization, culminating in a comprehensive, source-grounded research outline, where each section is explicitly linked via citations to a curated memory bank of source evidence. When it turns to the writing phase, to address the critical long context and attentional context management challenge, the writer executes a memory-grounded, citation-driven, hierarchical synthesis process. It constructs the report section by section, performing targeted retrieval of only the relevant evidence from a structured memory bank via citations in the outline for each subtask. This synergistic division of labor enables our agent to navigate complex information landscapes and produce reports that are both comprehensive in scope and trusted in their evidentiary grounding.

Extensive experiments demonstrate that WebWeaver achieves state-of-the-art (SOTA) performance and outperforms both the proprietary and open-source agent systems on three recent and challenging open-ended deep research benchmarks. Detailed discussion is produced to demonstrate the effectiveness of outline optimization and memory-grounded synthesis. Critically, WebWeaver enables agentic finetuning of small models for practical use. We construct a high-quality SFT dataset, WebWeaver-3k, generated by our framework. The experiments with WebWeaver-3k demonstrate that the complex skills of thinking, searching, and writing can be learned, enabling smaller, accessible models to achieve the expert-level performance previously confined to large-scale proprietary systems.

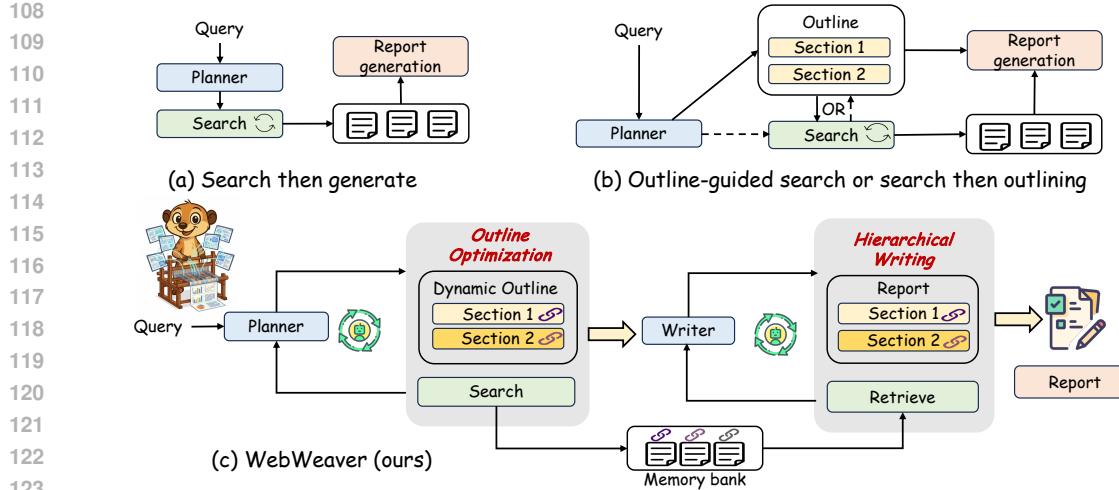


Figure 3: (a) the search-then-generate paradigm first gathers information and then directly generates a report; (b) the paradigms decouple the searching and outline generation; (c) WebWeaver not only enables a dynamic research cycle where the outline and search strategy co-evolve but allows hierarchical and attentional writing by retrieving only relevant evidence with citations in the outline.

2 PRELIMINARIES

Problem definition. We consider the open-ended research question without the ground-truth answers. Given an open-ended question, the agents need to search relevant information and finally output a report or article. To achieve this, we implement a planner for collecting information, a memory to store materials, and a writer for report generation. For both the planner and writer, we adopt ReAct (Yao et al., 2023) as the agent’s framework. Upon receiving a question, they perform several iterations of *thought-action-observation*. Specifically, in each iteration, based on the existing context, the LLM generates a *thought* and executes a parsable *action*, then awaits the environment to return an *observation*. The planning and writing stages terminate with the output token of “<terminate>”. A complete trajectory with T iterations can be defined as

$$\mathcal{H}_T = (\tau_0, a_0, o_0, \dots, \tau_i, a_i, o_i, \dots, \tau_T, a_T), \quad (1)$$

where τ_i , a_i , o_i represent thought, action, and observation sampled from the planning or writing policy based on all previous context in the i -th round, respectively.

Actions. For the planner, the action space consists of *search*, *write outline*, and *terminate*. Given the search queries, the search engine returns titles, snippets, and corresponding URLs. To save context space, we further execute the actions of URL selection, parsing pages via URLs, summarizing relevant contents, and extracting evidence with LLMs following the searching queries. The search tool finally returns the selected URLs with their corresponding summaries and evidence. The action of “*write outline*” is to generate and optimize the outline with citations linking to the evidence in the memory bank, and the “*terminate*” action is to terminate the planning process.

For the writer, the action space consists of *retrieve*, *write*, and *terminate*. Besides the *terminate* action, the *retrieve* action is to retrieve evidence from the memory bank by providing the grounded citations in the outline. The *write* action is provided to write the section of the report.

Memory bank. Answering an open-ended question requires long-context input of the collected information and long-context output of the final report. To search sufficient materials, the planner often searches and parses more than 100 web pages, with more than 100k tokens. The writer often outputs more than 20k tokens to produce a comprehensive report. Prior open-sourced deep research agents (Roucher et al., 2025; Research, 2025e;c) include all the raw materials (e.g., web pages and PDF files) in the LLM context, leading to quality degradation due to attentional failures like the “lost in the middle” problem, poor coherence, and increased hallucinations (Liu et al., 2023; Li et al., 2024a; Bai et al., 2024; Wu et al., 2025c). To this end, we introduce a memory to achieve context management for both planner and writer. Only a short summary of the web page or PDF file is

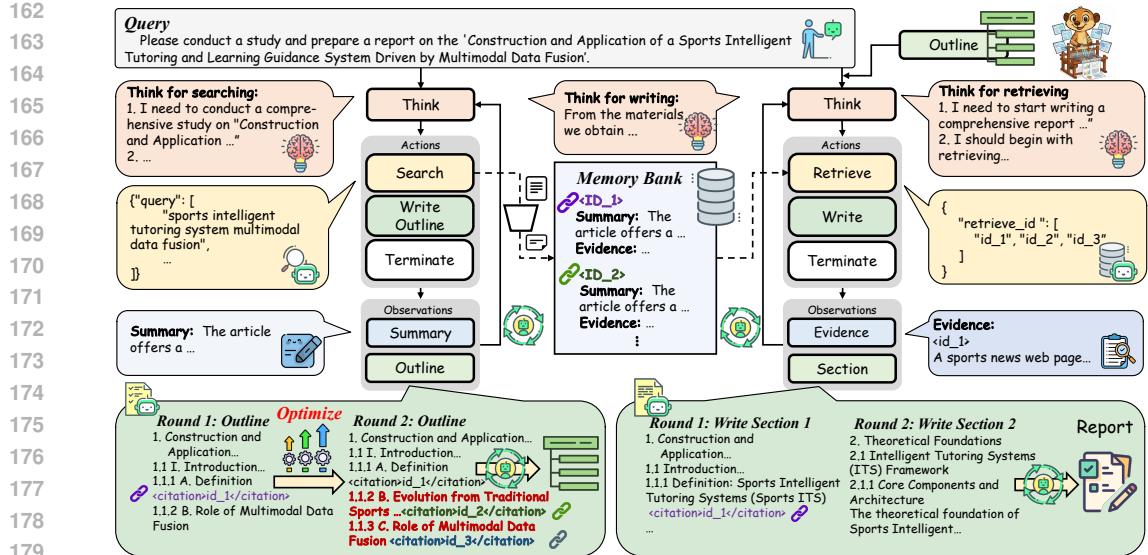


Figure 4: The workflow of WebWeaver. Left: The planner first iteratively collects evidence via the search action and optimizes the outline until outputting a comprehensive and citation-grounded outline. Right: The writer performs hierarchical and attentional writing by retrieving relevant evidence with the grounded citations in the outline.

included in the search context, and only necessary raw pages will be retrieved from the memory to write the corresponding sections via the citations in the outline.

3 METHOD

Our methodology is embodied in a dual-agent framework, comprising a planner and a writer. The planner is responsible for the dynamic cycle of evidence acquisition and outline optimization, while the writer performs evidence-grounded, section-by-section synthesis to construct the final report. This division of labor directly mirrors the cognitive workflow of a human researcher. The entire workflow is visualized in Fig. 4.

The planner operates in a dynamic, agentic research cycle. It iteratively interleaves evidence acquisition from web searches with the continuous refinement and optimization of a report outline. The output of this exploratory phase is not just a collection of sources but a comprehensive, well-structured outline where each section is explicitly linked via citations to a curated memory bank of source evidence. Subsequently, the writer takes over for the synthesis phase. To circumvent the pitfalls of brute-force generation and noisy-context issues, the writer adopts a section-wise and memory-grounded synthesis approach. For each section of the outline, it performs targeted retrieval of only the pertinent evidence from the memory bank using citations and composes the content. This division of labor ensures that the final report is not only coherent and well-organized but also deeply source-grounded, faithfully mirroring the rigor of human-led deep research.

3.1 RESEARCH CYCLE: ITERATIVE EVIDENCE ACQUISITION AND OUTLINE OPTIMIZATION

Recent deep research agents primarily follow two distinct, yet equally flawed, paradigms that decouple planning from discovery. The first, an “outline-guided search” approach (Han et al., 2025; Research, 2025e;c), generates a static outline before any evidence is gathered. This rigidly confines the research to the LLM’s internal, often outdated knowledge, making it blind to emergent insights. The second, a “search-then-outlining” approach (Shao et al., 2024; Xiong et al., 2025), performs an initial, broad search and then generates a fixed outline based on these preliminary findings. This approach permanently constrains the research scope to what was discovered in the initial, undirected search. Despite their different ordering, both paradigms share the same critical limitation: their one-way workflow prevents adaptation. In contrast, we propose a dynamic research cycle where the outline

216 and search strategy co-evolve, creating a feedback loop that allows the planner to continuously adapt
 217 and explore in response to new findings.
 218

219 The core of our planner’s operation is a dynamic research cycle that iteratively interleaves evidence
 220 acquisition with outline optimization. Unlike static approaches, our planner continuously adapts
 221 its strategy based on emergent findings. For each step, the planner selects one of the three actions:
 222 search, write outline, and terminate.
 223

224 **Evidence acquisition.** When the planner thinks that there is still insufficient evidence or knowledge
 225 to make a comprehensive outline to answer the open-ended question, it will continue collecting
 226 evidence by executing the search action. Given any search queries, the planner begins by querying a
 227 web search engine, which returns the results that contain the raw URLs with corresponding snippets
 228 and titles. To combat the contextual noise and processing overhead from raw URLs, it employs a
 229 two-stage filtering process. First, we prompt LLMs to select only the relevant URLs based on titles
 230 and snippets. Then, for each parsed page of the selected URLs, we perform two critical actions:
 231 leveraging LLMs to (1) distill a query-relevant summary, which is fed back into the planner’s context
 232 to inform subsequent search iterations, and (2) extract verifiable, detailed evidence (e.g., quotes, data
 233 points), which is stored in a structured memory bank for the subsequent writing.
 234

235 **Outline optimization.** After acquiring some evidence, the planner revisits the report’s outline. This is
 236 not a one-time generation step but a process of continuous refinement and optimization. The planner
 237 uses the newly acquired information to expand sections, add citations, or even restructure the entire
 238 outline to better reflect a comprehensive understanding of the topic. Crucially, it populates the outline
 239 with citations, mapping each section to the specific evidence IDs in the memory bank. This citation
 240 mechanism is vital for ensuring source-groundedness and supporting the subsequent hierarchical
 241 writing stage. This iterative loop continues until the planner outputs a terminate action with a tag
 242 “<terminate>” when the outline is sufficiently comprehensive and well-supported by evidence.
 243

244 3.2 MEMORY-GROUNDED SYNTHESIS: HIERARCHICAL RETRIEVAL AND WRITING

245 A pivotal challenge in generating long-form reports is not just information access but attentional
 246 management. The prevailing approaches of feeding all gathered evidence (Bai et al., 2025; Xiong et al.,
 247 2025) or redundant content from retrieval (Shao et al., 2024) saturate the model’s attentional capacity,
 248 leading to the increased hallucinations (Liu et al., 2023), where crucial details are overlooked, and
 249 “contextual bleeding” (Liu et al., 2025), where information from one section incorrectly influences
 250 the synthesis of another. We argue that a successful synthesis process must mirror human cognition
 251 by breaking down the complex task of long-context writing into manageable subtasks of attentional
 252 writing with only correct, relevant evidence. Therefore, we adopt a hierarchical, citation-grounded
 253 strategy, where the report is constructed sequentially, with the grounded retrieval of evidence using
 254 citations in the outline to enhance the writing correctness.
 255

256 Upon completion of the planning phase, the writer is provided with the structured, source-grounded
 257 outline and access to the evidence memory bank. The composition of each section is not a single,
 258 monolithic action but a deliberate, intra-sectional reasoning cycle designed to ensure both accuracy
 259 and coherence. This cycle unfolds as follows:
 260

261 First, the writer identifies its immediate subtask, such as “Let’s write the first section.” It then executes
 262 a targeted retrieval action, pulling only the relevant evidence from the memory bank as indicated
 263 by the outline’s citations. With evidence, the writer enters a crucial reasoning phase with a think
 264 action. It analyzes the previous sections and retrieved content, synthesizes key insights, selects the
 265 most compelling pieces of evidence, and formulates a coherent narrative structure for the section.
 266 This internal monologue is critical for moving beyond simple summarization to genuine synthesis.
 267

268 Only after this internal analysis is formed does the writer proceed to the writing action, composing
 269 the prose and encapsulating it within “<write>” tags. Once a section is complete, its corresponding
 270 source materials are explicitly pruned from the context window and replaced with a placeholder
 271 message. This dynamic retrieval-and-pruning mechanism is the cornerstone of our approach: it
 272 ensures the writer’s context remains highly relevant and coherent for the next cycle, mitigates context
 273 overflow, and prevents cross-sectional interference. This entire process repeats hierarchically for all
 274 sections until the writer outputs “<terminate>” token, signaling the completion of the full report.
 275

Agent systems	RACE					FACT	
	Overall	Comp.	Insight	Inst.	Read.	Eff. c.	C. acc.
OpenScholar	25.56	21.38	15.49	38.58	29.77	-	-
WebShaper (32B)	34.93	31.58	26.17	44.81	40.38	-	-
ReAct (Qwen3-256b-a22b-instruct-2507)	46.16	45.04	43.2	49.16	46.17	-	-
langchain-open-deep-research	43.44	42.97	39.17	48.09	45.22		
doubaoresearch	44.34	44.84	40.56	47.95	44.69	52.62	52.86
kimi-research	44.64	44.96	41.97	47.14	45.59	-	-
Claude-research	45.00	45.34	42.79	47.58	44.66	-	-
openai-deepresearch	46.45	46.46	43.73	49.39	47.22	39.79	75.01
Gemini-2.5-pro-deepresearch	49.71	49.51	49.45	50.12	50.00	165.34	78.30
WebWeaver (Qwen3-30b-a3b-instruct-2507)	45.78 (0.7)	44.44 (0.5)	45.10 (0.5)	48.16 (0.8)	45.32 (1.2)	22.34 (2.6)	25.29 (0.9)
WebWeaver (GPT-oss-120b)	47.71 (0.8)	47.38 (0.3)	46.86 (0.1)	48.01 (0.5)	48.85 (0.5)	66.83 (1.5)	67.03 (0.5)
WebWeaver (Qwen3-235b-a22b-instruct-2507)	50.80 (0.1)	51.45 (0.1)	51.39 (0.2)	50.26 (0.1)	48.98 (0.2)	152.70 (6.3)	75.72 (3.1)
WebWeaver (Claude-sonnet-4-20250514)	50.48 (0.1)	51.65 (0.1)	49.67 (0.3)	50.57 (0.1)	49.72 (0.1)	216.99 (6.8)	92.13 (0.7)

Table 1: Performance of agents on DeepResearch Bench in terms of comprehensiveness (Comp.), insight, instruction-following (Inst.), readability (Read.), effective citations (Eff. c.), and citation accuracy (C. acc.). The best results are highlighted with purple color, and the second-best results are highlighted with underlines. Standard deviations are present in bracket.

4 EXPERIMENTS

In this section, we first evaluate WebWeaver on three recent and challenging benchmarks. Detailed discussion is then produced to demonstrate the effectiveness of outline optimization and memory-grounded synthesis. Furthermore, we curate a high-quality SFT dataset, WebWeaver-3k, to improve the capabilities of thinking, searching, and writing for a smaller model to achieve expert-level performance, which is provided in Appendix C.

4.1 EXPERIMENTAL SETUPS

Benchmarks. To evaluate the performance of Deep Research systems, we use three open-ended benchmark datasets: (1) **DeepResearch Bench** (Du et al., 2025) comprises 100 PhD-level complex research tasks meticulously formulated by domain experts across 22 distinct fields, such as Science & Technology, Finance & Business, Software Engineering, and Art & Design. (2) **DeepConsult** (Consult, 2025) is a specialized collection of prompts tailored for in-depth research within the business and consulting domains. The query set encompasses a wide range of topics, including marketing strategy, financial analysis, emerging technology trends, and business planning. (3) **DeepResearchGym** (Coelho et al., 2025) is used to assess performance on real-world, complex queries. This dataset contains 100 queries sampled from the extensive Researchy Questions dataset (Rosset et al., 2024), which includes approximately 96,000 authentic information-seeking queries. The metrics of the benchmarks are detailed in Appendix B.

Compared systems. We benchmark the performance of WebWeaver against a range of state-of-the-art DeepResearch systems. These systems are categorized into two groups: (1) Open-Source Systems: For open-source counterparts, we compare against OpenScholar (Asai et al., 2024), WebShaper-32B (Tao et al., 2025), and langchain-open-deep-research (LangChain, Inc., 2023). We also include the widely-used ReAct paradigm (Yao et al., 2023) equipped with the identical backbone LLM (Qwen-235b-a30b-instruct-2507) and the same tools; (2) The Powerest Proprietary Systems from DeepResearch Bench leaderboard¹ (up to Sept. 25, 2025): doubaoresearch (Research, 2025a), kimi-research (Research, 2025d), Claude-research (anthropic, 2025), openai-deepresearch (OpenAI, 2025a), and Gemini-2.5-pro-deepresearch (Research, 2025b).

Implementation details. The WebWeaver is compatible with various advanced LLMs. In the experiments, we utilize the following models: Qwen3-30b-a3b-instruct-2507 (Yang et al., 2025), GPT-oss-120b (Agarwal et al., 2025), Qwen3-235b-a22b-instruct-2507 (Yang et al., 2025), and Claude-sonnet-4-20250514 (anthropic, 2025). We adopt Claude-sonnet-4-20250514 as the default agent model for ablation studies and discussion without any statements. We use GPT-oss-120b to select relevant URLs, perform query-relevant summaries, and extract evidence for the search action.

¹<https://huggingface.co/spaces/Ayanami0730/DeepResearch-Leaderboard>

324

325

Agent systems	DeepConsult				DeepResearchGym						
	win	tie	lose	Avg. score	Cl.	Depth	Bal.	Brea.	Sup.	Ins.	Avg. score
WebShaper (32B)	3.25	3.75	93.00	1.63	64.70	63.00	59.30	66.50	9.40	59.90	53.80
<u>ReAct (Qwen3-235b-a22b-instruct-2507)</u>	<u>51.55</u>	<u>38.66</u>	<u>9.79</u>	<u>5.83</u>	<u>89.49</u>	<u>95.45</u>	<u>88.08</u>	<u>95.66</u>	<u>62.32</u>	<u>89.29</u>	<u>86.72</u>
doubaoresearch	29.95	40.35	29.70	5.42	68.85	93.12	83.96	93.33	84.38	83.12	84.46
Claude-research	25.00	38.89	36.11	4.60	86.67	96.88	84.41	96.56	26.77	90.22	80.25
openai-depresearch	0.00	100.00	0.00	5.00	84.90	98.10	89.80	97.40	88.40	89.00	91.27
Gemini-2.5-pro-depresearch	61.27	31.13	7.60	6.70	90.71	99.90	93.37	99.69	95.00	97.45	96.02
WebWeaver (Qwen3-30b-a3b-instruct-2507)	23.79 (2.7)	28.64 (3.2)	47.57 (5.9)	3.99 (0.3)	72.48 (0.3)	85.32 (0.5)	75.78 (0.4)	84.25 (0.3)	65.86 (0.9)	81.97 (0.2)	77.61 (0.2)
WebWeaver (gpt-oss-120b)	57.18 (3.7)	22.94 (4.6)	19.89 (1.4)	6.42 (0.1)	87.89 (1.5)	96.62 (0.6)	88.12 (0.6)	95.77 (0.5)	91.69 (3.1)	93.57 (0.3)	92.28 (1.0)
<u>WebWeaver (Qwen3-235b-a22b-instruct-2507)</u>	67.69 (1.1)	15.94 (4.2)	16.37 (4.4)	6.90 (0.1)	89.71 (0.3)	99.95 (0.1)	91.95 (0.3)	99.75 (0.1)	96.70 (1.2)	95.65 (0.5)	95.57 (0.4)
WebWeaver (Claude-sonnet-4-20250514)	66.16 (1.3)	12.14 (2.0)	21.68 (0.7)	6.94 (0.1)	90.37 (0.1)	99.96 (0.1)	94.12 (0.2)	99.90 (0.1)	98.69 (0.2)	97.37 (0.2)	96.74 (0.1)

328

Table 2: Performance of agents on DeepConsult in terms of win rate and average scores and on DeepResearchGym in terms of clarity (Cl.), depth, balance (Bal.), breadth (Brea.), support (Sup.), and insightfulness (Ins.). The best results are highlighted with purple color, and the second-best results are highlighted with underlines. Standard deviations are present in bracket.

331

We run three independent runs and report means and standard deviations. The detailed cost analysis is discussed in Appendix F. We present the prompts and case studies in Appendix I.

339

4.2 MAIN RESULTS

340

Results on DeepResearch Bench. As presented in Table 1, our WebWeaver framework establishes a new state-of-the-art, consistently outperforming existing agents. This superior performance is a direct result of our dual-agent, iterative methodology. The high scores in comprehensiveness (Comp.) and insight stem from the planner’s dynamic research cycle, which iteratively expands the report’s scope based on emergent findings. This process naturally leads to a higher number of effective citations (Eff. c.), as the planner is intrinsically motivated to seek more evidence to ensure that each section is well-supported. Furthermore, the remarkable citation accuracy (C. acc.) of 92.13% is achieved by the strong synergy between our agents: the planner embeds specific citation IDs into the outline, and the writer’s hierarchical synthesis process uses this structure for targeted retrieval. By focusing only on relevant evidence for each section, it drastically reduces context-bleeding and hallucinations, which also contributes to the enhanced readability (Read.), instruction-following (Inst.) scores, and citation accuracy (C. acc.). This demonstrates that by emulating human research patterns, our framework produces not just more thorough but also significantly more reliable and well-structured reports.

354

Results on DeepConsult and DeepResearchGym. To validate its generalizability, we evaluated WebWeaver on DeepConsult and DeepResearchGym (Table 2), where it achieved the highest win rate (67.69%) and top average score (96.74), respectively. This success is rooted in our core design: the planner’s iterative cycle ensures near-perfect depth and breadth, while the writer’s hierarchical synthesis guarantees outstanding balance and support. These quantitative strengths in structural metrics directly explain the qualitative victories, proving the robustness of our writing strategy. A detailed human evaluation is present in Appendix E to further demonstrate WebWeaver’s superiority in report generation.

362

4.3 ANALYSIS

363

Statistics of planning and writing. The statistics in Table 3 provide a compelling quantitative narrative that not only justifies but also demonstrates the benefits of WebWeaver’s design. The planning task involves an extensive exploratory phase with nearly 16 search steps and 21 unique search queries, proving that a simple, linear search is insufficient. The critical finding is that the outline undergoes more than two optimization cycles on average, expanding into a complex 4k-token outline. This empirically invalidates static-outline approaches and shows the tangible benefit of our iterative process: it produces a richer, more comprehensive plan that adapts to discovery. This deep planning phase amasses a staggering amount of information—over 100 saved pages, culminating in 67k evidence

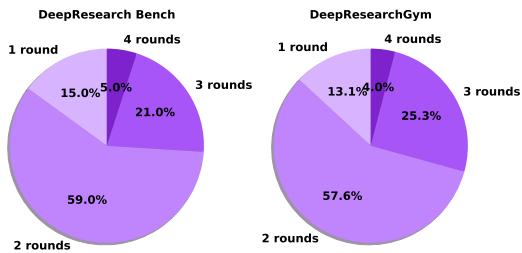


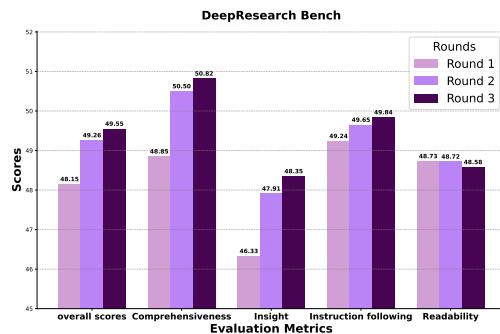
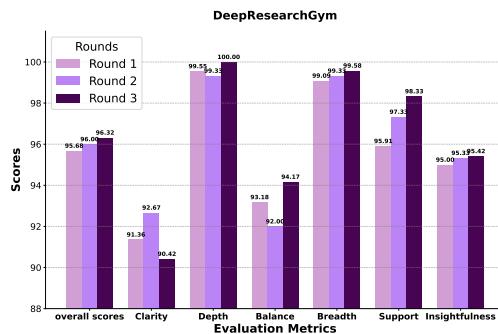
Figure 5: Statistics of outline optimization of Claude-sonnet-4-20250514 on DeepResearch Bench and DeepResearchGym.

	Planning statistics							Writing statistics	
	# Search step	# Outline token	# Outline optimization	# Saved page	# Search query	# Evidence token	# Summary token	# Output token	# Writing step
DeepResearch Bench	15.71	4876.21	2.16	112.25	20.24	67237	14980	26127	24.78
DeepResearchGym	16.65	3732.87	2.20	102.55	21.93	66301	12543	26004	24.71

383 Table 3: The planning and writing statistics of Claude-sonnet-4-20250514 on benchmarks.

386 tokens and 15k summary tokens. This sheer volume makes a brute-force approach computationally
387 hard, thus mandating our memory-centric architecture with targeted retrieval as a foundational
388 requirement, not just an optimization. Finally, the writer’s process of composing a 26k-token report
389 in 25 discrete writing steps validates that our hierarchical synthesis is a practical way to maintain
390 coherence over long outputs. In essence, the statistics of searching and writing affirm that each
391 component of WebWeaver is a necessary and beneficial response to the inherent challenges of OEDR.

392 **Superiority of outline optimization.** To isolate and quantify the benefits of outline optimization,
393 as reported in Fig. 6, 7, we conducted an ablation study by evaluating the end-to-end benchmark
394 performance. We collect the samples with three-round outline optimization from DeepResearch
395 Bench and DeepResearchGym, adopting the same writing strategy for them. The first-round optimized
396 outline scenario does not guide searching here, which is similar to the planning paradigm of the prior
397 work STORM (Shao et al., 2024).

409 Figure 6: End-to-end scores with varying rounds
410 of outline optimization on DeepResearch Bench.509 Figure 7: End-to-end scores with varying rounds
510 of outline optimization on DeepResearchGym.

512 The benefits of this iterative refinement are evident across both benchmarks. On DeepResearch Bench,
513 the overall score steadily climbs, driven primarily by significant gains in comprehensiveness (48.85
514 → 50.82) and insight (46.33 → 48.35). This directly validates our hypothesis that each optimization
515 round allows the planner to build a more detailed and logically structured outline. This enhanced
516 structure is further reflected in DeepResearchGym’s metrics, where later rounds achieve near-perfect
517 scores in depth (100) and breadth (99.58), indicating a more exhaustive topic coverage. Crucially,
518 this is not just about adding more content; the steady rise in support (95.91 → 98.33) demonstrates
519 that a more refined outline creates a better-scaffolded structure, enabling the writer to more tightly
520 link claims to evidence. In summary, this analysis empirically demonstrates that iterative outline
521 optimization is not a redundant step but a critical mechanism for elevating a report from a simple
522 summary to a deep, insightful, and well-supported piece of research. We also provide LLM as a judge
523 (Zheng et al., 2023) for multiple-round optimized outlines in Appendix G.

524 **Hierarchical retrieval and writing vs. brute-force writing.** To empirically validate our hierarchical
525 writing process, we conducted a critical ablation study comparing our hierarchical writer against
526 a brute-force baseline that attempts to include the entire memory bank to generate the final report
527 in multiple steps, which is similar to the workflow of LongWriter (Bai et al., 2025). **No evidence**
528 **chunking here without exceeding the context limits of Claude-sonnet-4-20250514.** The results in
529 Fig. 8, 9 are unequivocal: our hierarchical approach dramatically outperforms the brute-force method
530 across every metric, confirming that a hierarchical writing strategy is essential. The most striking
531 improvements are in insight (42.72 → 49.47) and readability (43.82 → 49.72), which directly
532 validates our hypothesis on attentional management; by focusing the model on a curated context for

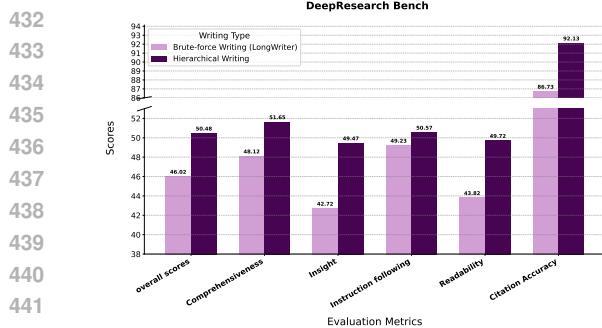


Figure 8: Performance comparison between hierarchical writing and brute-force writing (LongWriter) on DeepResearch Bench.

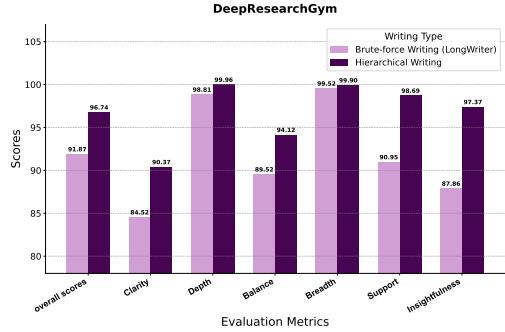


Figure 9: Performance comparison between hierarchical writing and brute-force writing (LongWriter) on DeepResearchGym.

each section, it can perform deeper reasoning rather than shallow summarization. Furthermore, the accurate retrieved content using citations also brings higher citation accuracy ($86.73 \rightarrow 92.13$) and supportiveness ($90.95 \rightarrow 98.69$). In conclusion, these results provide definitive evidence that emulating the human cognitive process of focused, section-by-section writing is not merely a beneficial choice but a fundamental requirement for generating coherent, insightful, and reliable long-form reports.

Context management superiority. The context management superiority of our hierarchical writing process is quantitatively validated in Figure 10. At each individual writing step, our method provides the model with a significantly smaller, more focused context compared to the brute-force baseline. This design dramatically reduces the model’s cognitive load. Freed from the burden of navigating a cluttered and noisy context, the writer can perform more thorough synthesis for each section. This leads to a richer output that cumulatively grows with each step, and as the plot clearly indicates, surpasses the baseline in total output tokens precisely at step 6, ultimately generating a substantially longer and more detailed final report. This demonstrates that the key to producing extensive and well-developed content lies not in overwhelming the model with information, but in providing the right information at the right time.

5 RELATED WORKS

Open-Ended Deep Research. Deep Research Agents have garnered significant attention for their powerful capabilities in information seeking, integration, and reasoning. Proprietary systems, such as DeepResearch (OpenAI, 2025a), Gemini Deep Research (google, 2025), and Claude Research (anthropic, 2025), have demonstrated performance comparable to human experts in domains like fact-checking and report writing. However, their opaque internal architectures and workflows hinder broader research and development. In the open-source community, many studies (Li et al., 2025b; Tao et al., 2025; Su et al., 2025; Qiao et al., 2025; Fang et al., 2025; Li et al., 2025a; Wu et al., 2025b;a; Li et al., 2024b) have been developed to tackle complex research Question-Answering (QA) benchmarks. Nevertheless, these solutions are primarily tailored for short-answer research queries and lack the capability to generate comprehensive, long-form reports on open-domain topics. Other open-source systems like OpenDeepResearch (Research, 2025e), GPT Researcher (Research, 2025c), and TTD-DR (Han et al., 2025) address long-form generation by first drafting a static framework, then retrieving content, and finally composing the report. This approach, characterized by a fixed structure and one-step generation, often leads to textual incoherence and hallucinations. While recent works like WriteHere (Xiong et al., 2025), STORM (Shao et al., 2024), and SCISAGE (Shi et al., 2025) utilize searched content to generate or refine the outline, the search-first method also constrains

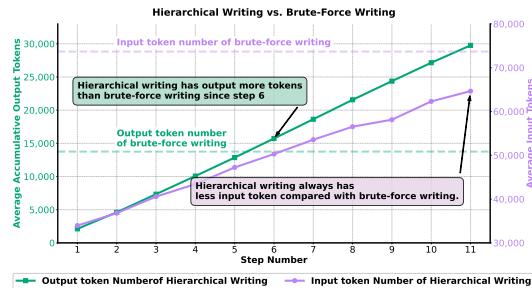


Figure 10: Input and output token statistics of Claude-sonnet-4-20250514 for hierarchical and brute-force writing on DeepResearch Bench.

486 the searching boundary without outline guidance. In sharp contrast, WebWeaver enables a truly
 487 synergistic research cycle where the outline and search strategy co-evolve, allowing emergent findings
 488 to continuously reshape the research direction in real time.

489 **Long Writing.** Ensuring the coherence and accuracy of LLM-generated long-form text is a persistent
 490 challenge. Previous work has explored methods like recursive prompting for story extension (Yang
 491 et al., 2022) and structured task decomposition to improve consistency (Yang et al., 2023; Wang
 492 et al., 2025; Huot et al., 2025). More recently, agent-based frameworks have become a mainstream
 493 solution. Systems like LongWriter (Bai et al., 2025), and CogWriter (Wan et al., 2025) employ a
 494 "plan-then-write" strategy, where a planner first creates an outline, and a writer then conditions on
 495 this plan to produce the full text. However, these methods rely on a static initial plan and a brute-force
 496 writing strategy by feeding all the evidence into LLMs. In contrast, our approach uniquely enables
 497 the outline to be dynamically optimized in tandem with the evidence acquisition process, allowing
 498 for a comprehensive, source-grounded research outline. While recent works (Huot et al., 2025; Shao
 499 et al., 2024) adopted the multi-agent paradigm to write sections in parallel, with evidence retrieval
 500 based on the section title, the separate writing often leads to content and style incoherence, and the
 501 retrieval using titles also brings noisy evidence. In contrast, our agentic hierarchical writing model is
 502 designed to foster both global coherence and local depth. Its sequential, single-agent process allows
 503 cross-sectional thinking by maintaining a continuous narrative flow between sections. At the same
 504 time, its reliance on precise, citation-grounded evidence enables deep internal reasoning within each
 505 section. This combination allows the writer to produce a cohesive, accurate, and insightful report.

507 6 CONCLUSION

508 In this paper, we introduced WebWeaver, a novel dual-agent framework designed to overcome
 509 the fundamental flaws of static, machine-like pipelines in open-ended deep research (OEDR). By
 510 emulating the human cognitive process that integrates the planner’s dynamic research cycle with
 511 the writer’s hierarchical retrieval and writing process, WebWeaver consistently outperforms both
 512 proprietary and open-source systems, establishing a new state-of-the-art.

513 Beyond its superior performance, the true significance of WebWeaver lies in the paradigm for tackling
 514 information-intensive tasks. It demonstrates that intractable long-context input and output can be
 515 deconstructed into a solvable problem of system-level context management, orchestrated through
 516 deliberate, tool-driven actions. This work, therefore, does not just present better agents but offers a
 517 new system blueprint to master knowledge through purposeful action, not brute-force attention.

521 ETHICS STATEMENT

522 This work adheres to the ICLR Code of Ethics. The WebWeaver agent is designed to be a responsible
 523 web citizen. It programmatically respects robots.txt directives, and all web access is conducted for
 524 non-commercial academic research, aligning with fair use principles often permitted under website
 525 terms of service. Furthermore, we have implemented safeguards regarding Personally Identifiable
 526 Information (PII) at multiple stages. Our dataset curation process includes a mandatory step to
 527 anonymize personal information. During live evidence extraction, the agent’s prompts are designed to
 528 focus on factual evidence about the queries. Beyond these specific measures, we have taken steps
 529 to identify and mitigate potential biases in data selection, model prompting, and system design to
 530 ensure fairness and avoid harmful outcomes. The system is developed solely for research purposes,
 531 with these safeguards in place to minimize misuse. No human subjects or animal experiments were
 532 involved in this study.

535 536 REPRODUCIBILITY STATEMENT

537 To ensure the reproducibility of our work, we provide detailed information regarding the training
 538 details, prompts, and case studies. Appendix C of this paper includes the training details for our
 539 agentic finetuning. The prompts and case studies of planner and writer are detailed in Appendix I.

540 REFERENCES
541

542 Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K
543 Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv*
544 *preprint arXiv:2508.10925*, 2025.

545 anthropic. Meet claude, 2025. URL <https://www.anthropic.com/claude>.
546

547 Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo,
548 Luca Soldaini, Sergey Feldman, Mike D’arcy, et al. Openscholar: Synthesizing scientific literature
549 with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.

550

551 Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi
552 Li. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint*
553 *arXiv:2408.07055*, 2024.

554 Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi
555 Li. Longwriter: Unleashing 10,000+ word generation from long context llms. In *The Thirteenth*
556 *International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
557 OpenReview.net, 2025. URL <https://openreview.net/forum?id=kQ5s9Yh0WI>.
558

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

562 Deep Consult. Deep consult. 2025. URL <https://github.com/Su-Sea/ydc-deep-researc-h-evals>.
563

564 Google DeepMind. Gemini 2.5, 2025. URL <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>.
565

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

568 Runnan Fang, Shihao Cai, Baixuan Li, Jialong Wu, Guangyu Li, Wenbiao Yin, Xinyu Wang, Xiaobin
569 Wang, Liangcai Su, Zhen Zhang, Shibin Wu, Zhengwei Tao, Yong Jiang, Pengjun Xie, Fei Huang,
570 and Jingren Zhou. Towards general agentic intelligence via environment scaling, 2025.

571

572 google. Try deep research and our new experimental model in gemini, your ai assistant, 2025. URL
573 <https://blog.google/products/gemini/google-gemini-deep-researc-h/>.
574

575 Rujun Han, Yanfei Chen, Zoey CuiZhu, Lesly Miculicich, Guan Sun, Yuanjun Bi, Weiming Wen,
576 Hui Wan, Chunfeng Wen, Solène Maître, George Lee, Vishy Tirumalashetty, Emily Xue, Zizhao
577 Zhang, Salem Haykal, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. Deep researcher with
578 test-time diffusion. *CoRR*, abs/2507.16075, 2025. doi: 10.48550/ARXIV.2507.16075. URL
579 <https://doi.org/10.48550/arXiv.2507.16075>.
580

581 Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth
582 Clark, and Mirella Lapata. Agents’ room: Narrative generation through multi-step collaboration.
583 In *The Thirteenth International Conference on Learning Representations*, 2025.

584

585 Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language
586 models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

587

588 LangChain, Inc. LangChain: Building applications with LLMs through composability, 2023. URL
589 <https://python.langchain.com/>.
590

591 Kuan Li, Zhongwang Zhang, Hufeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu
592 Zhang, Xixi Wu, Jialong Wu, Xinyu Wang, Zile Qiao, et al. Websailor-v2: Bridging the chasm to
593 proprietary agents via synthetic data and scalable reinforcement learning, 2025a.

594 Kuan Li, Zhongwang Zhang, Hufeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan
 595 Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong
 596 Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating super-human
 597 reasoning for web agent. *CoRR*, abs/2507.02592, 2025b. doi: 10.48550/ARXIV.2507.02592.
 598 URL <https://doi.org/10.48550/arXiv.2507.02592>.

599 Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with
 600 long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024a.

601

602 Zhonghao Li, Xuming Hu, Aiwei Liu, Kening Zheng, Sirui Huang, and Hui Xiong. Refiner:
 603 Restructure retrieval content efficiently to advance question-answering capabilities. *arXiv preprint*
 604 *arXiv:2406.11357*, 2024b.

605 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 606 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint*
 607 *arXiv:2412.19437*, 2024.

608

609 Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni,
 610 and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint*
 611 *arXiv:2307.03172*, 2023.

612 Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John,
 613 Yassine Benajiba, and Dan Roth. Towards long context hallucination detection. *arXiv preprint*
 614 *arXiv:2504.19457*, 2025.

615 Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia:
 616 a benchmark for general ai assistants. In *The Twelfth International Conference on Learning*
 617 *Representations*, 2023.

618

619 OpenAI. Deep research system card, 2025a. URL <https://cdn.openai.com/deep-resea>
 620 *rch-system-card.pdf*.

621 OpenAI. Introducing openai o3 and o4-mini, 2025b. URL <https://openai.com/index/i>
 622 *ntroducing-o3-and-o4-mini/*.

623

624 Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang,
 625 Baixuan Li, Hufeng Yin, Kuan Li, Rui Min, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang,
 626 and Jingren Zhou. WebResearcher: Unleashing unbounded reasoning capability in Long-Horizon
 627 Agents, 2025.

628

629 Qwen Team. QwQ-32B: Embracing the power of reinforcement learning, March 2025. URL
<https://qwenlm.github.io/blog/qwq-32b/>.

630

631 Doubao Deep Research. Doubao deep research. 2025a. URL <https://www.doubao.com/c>
 632 *hat/*.

633

634 Gemini Research. Gemini research. 2025b. URL <https://gemini.google/overview/d>
 635 *eep-research/*.

636

637 GPT Research. Gpt research. 2025c. URL <https://github.com/assafelovic/gpt-r>
 638 *esearcher*.

639

640 Kimi Deep Research. Kimi deep research. 2025d. URL <https://www.kimi.com/>.

641

642 Open Deep Research. Open deep research. 2025e. URL <https://github.com/langchain>
 643 *-ai/open_deep_research*.

644

645 Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C. Chau, Zhuo Feng, Ahmed Awadallah,
 646 Jennifer Neville, and Nikhil Rao. Researchy questions: A dataset of multi-perspective, decomposi-
 647 tional questions for LLM web agents. *CoRR*, abs/2402.17896, 2024. doi: 10.48550/ARXIV.2402.
 648 17896. URL <https://doi.org/10.48550/arXiv.2402.17896>.

649

650 Aymeric Roucher, Albert Villanova del Moral, merve, Thomas Wolf, and Clémentine Fourrier. Open-
 651 source deepresearch – freeing our search agents. 2025. URL <https://huggingface.co/b>
 652 *log/open-deep-research*.

648 Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting
 649 in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the*
 650 *2024 Conference of the North American Chapter of the Association for Computational Linguistics:*
 651 *Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024.

652 Xiaofeng Shi, Qian Kou, Yuduo Li, Ning Tang, Jinxin Xie, Longbin Yu, Songjing Wang, and Hua
 653 Zhou. Scisage: A multi-agent framework for high-quality scientific survey generation. *arXiv*
 654 *preprint arXiv:2506.12689*, 2025.

655 Liangcai Su, Zhen Zhang, Guangyu Li, Zhuo Chen, Chenxi Wang, Maojia Song, Xinyu Wang, Kuan
 656 Li, Jialong Wu, Xuanzhong Chen, Zile Qiao, Zhongwang Zhang, Huifeng Yin, Shihao Cai, Runnan
 657 Fang, Zhengwei Tao, Wenbiao Yin, et al. Scaling agents via continual pre-training, 2025.

658 Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li,
 659 Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webshaper:
 660 Agentically data synthesizing via information-seeking formalization. *CoRR*, abs/2507.15061, 2025.
 661 doi: 10.48550/ARXIV.2507.15061. URL <https://doi.org/10.48550/arXiv.2507.15061>.

662 Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. A cognitive writing
 663 perspective for constrained long-form text generation. In Wanxiang Che, Joyce Nabende, Ekaterina
 664 Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational*
 665 *Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 9832–9844. Association for
 666 Computational Linguistics, 2025. URL <https://aclanthology.org/2025.findings-acl.511/>.

667 Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan.
 668 Generating long-form story using dynamic hierarchical outlining with memory-enhancement.
 669 In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of*
 670 *the Nations of the Americas Chapter of the Association for Computational Linguistics: Human*
 671 *Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA,*
 672 *April 29 - May 4, 2025*, pp. 1352–1391. Association for Computational Linguistics, 2025. doi:
 673 10.18653/V1/2025.NAACL-LONG.63. URL <https://doi.org/10.18653/v1/2025.naacl-long.63>.

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

677 Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang,
 678 Zekun Xi, Yong Jiang, Pengjun Xie, et al. Webdancer: Towards autonomous information seeking
 679 agency. *arXiv preprint arXiv:2505.22648*, 2025a.

680 Weiqi Wu, Xin Guan, Shen Huang, Yong Jiang, Pengjun Xie, Fei Huang, Jiuxin Cao, Hai Zhao,
 681 and Jingren Zhou. Masksearch: A universal pre-training framework to enhance agentic search
 682 capability. 2025b. URL <https://arxiv.org/abs/2505.20285>.

683 Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. Longgenbench: Benchmarking
 684 long-form generation in long context llms. In *The Thirteenth International Conference on Learning*
 685 *Representations*, 2025c.

686 Ruibin Xiong, Yimeng Chen, Dmitrii Khizbulin, Mingchen Zhuge, and Jürgen Schmidhuber. Beyond
 687 outlining: Heterogeneous recursive planning for adaptive long-form writing with language models.
 688 *arXiv preprint arXiv:2503.08275*, 2025.

689 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
 690 Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*,
 691 2025.

692 Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with
 693 recursive reprompting and revision. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.),
 694 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*,

702 *EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 4393–4479. Association
 703 for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.296. URL
 704 <https://doi.org/10.18653/v1/2022.emnlp-main.296>.

705 Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. DOC: improving long story coherence
 706 with detailed outline control. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.),
 707 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume
 708 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 3378–3465. Association for
 709 Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.190. URL <https://doi.org/10.18653/v1/2023.acl-long.190>.

710 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.
 711 React: Synergizing reasoning and acting in language models. In *International Conference on
 712 Learning Representations (ICLR)*, 2023.

713 Haopeng Zhang, Philip S Yu, and Jiawei Zhang. A systematic survey of text summarization: From
 714 statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41, 2025.

715 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 716 Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.
 717 Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.

718
 719
 720
 721
 722
 723
 724
 725
 726
 727
 728
 729
 730
 731
 732
 733
 734
 735
 736
 737
 738
 739
 740
 741
 742
 743
 744
 745
 746
 747
 748
 749
 750
 751
 752
 753
 754
 755

756 Appendix

759 A USE OF LARGE LANGUAGE MODELS

760 In adherence to ICLR 2026 policy, we disclose the assistive use of Large Language Models (LLMs)
 761 in the preparation of this paper. The models were employed to refine grammar and improve the
 762 clarity of the text. Furthermore, LLMs function as the agents for the open-ended deep research. The
 763 authors have reviewed all LLM-generated contributions and take full responsibility for the content
 764 and integrity of this work.

766 B EXPERIMENT DETAILS

768 **Search tool.** WebWeaver accesses the Google search engine for information retrieval. It allows
 769 searching multiple queries simultaneously and returns the top-10 results for each query. Each
 770 result contains a title, a snippet, and the corresponding URL. To parse the web pages and pdf files,
 771 ScraperAPI² is first used to retrieve their full content, and then a summary model extracts relevant
 772 information based on the goal. In this paper, we use GPT-oss-120b as the summary model.

774 **Metrics.** We use the official evaluation metrics with the recommended judge LLMs of each
 775 benchmark. (1) **DeepResearch Bench.** This benchmark utilizes two suites of metrics to evaluate
 776 different aspects of the system’s output: **1) RACE (Report Quality):** It assesses the quality of the
 777 final generated report with a reference report across four dimensions, namely Comprehensiveness
 778 (Comp.), Insight/Depth (Insight), Instruction-Following (Inst.), and Readability (Read.). An overall
 779 score is then calculated as a weighted summation of these components. **2) FACT (Web Retrieval):** It
 780 measures the effectiveness and reliability of the information retrieval process. This includes Citation
 781 Accuracy (C. Acc.) and the Average Effective Citations per Task (Eff. c.). We adopt Gemini-2.5-
 782 pro as the judgement model by following the benchmark. (2) **DeepConsult.** Performance on this
 783 benchmark is determined through a pairwise comparison against the openai-deepsearch baseline.
 784 The primary metrics are the win rate, tie rate, and loss rate, which are supplemented by a reported
 785 average quality score. The judgement model is gpt-4.1-20250414. (3) **DeepResearchGym.** An
 786 LLM acts as a judge to assess the generated report on several quality dimensions, including clarity,
 787 insightfulness, depth, balance, breadth, support, and an average quality score. The judgement model
 788 is gpt-4.1-mini-20250414.

	Planning statistics							Writing statistics	
	# Search step	# Outline token	# Outline optimization	# Saved page	# Search query	# Evidence token	# Summary token	# Output token	# Writing step
WebWeaver-3k	14.67	4148.57	2.18	106.65	18.8	62637	14155	22637	22.76

793 Table 4: The planning and writing statistics of training data on WebWeaver-3k.

795 C AGENTIC FINETUNING

797 While 30B-scale LLMs (e.g., Qwen3-30b-a3b-instruct-2507) possess strong foundational capabilities,
 798 they often exhibit deficiencies in stability and instruction-following when executing complex, multi-
 799 turn tool-calling sequences over long contexts. To bridge this critical gap, we constructed a high-
 800 quality Supervised Fine-Tuning (SFT) dataset: WebWeaver-3k. The process began by sourcing a
 801 diverse set of queries crawled from the web.

802 We crawled web pages from a curated list of high-quality English and Chinese forums and communities,
 803 specifically targeting knowledge-intensive sources like Reddit. Our query generation process
 804 involved two main stages: We first used pattern matching to extract a large pool of potential queries.
 805 This included text from specific HTML tags (e.g., `<h1>`, `<h2>`, post titles) and sentences ending
 806 with a question mark, resulting in an initial set of over candidates. We then utilized LLM to filter and
 807 refine these candidates into high-quality, open-ended questions. Each candidate was evaluated using
 808 a carefully designed prompt that instructed the model to assess it against several criteria:

809 ²<https://www.scraperaapi.com/>

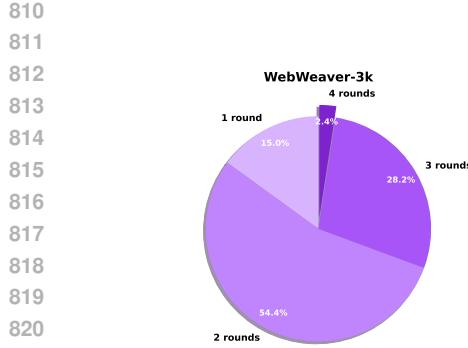


Figure 11: Round statistics of outline optimization on WebWeaver-3k.

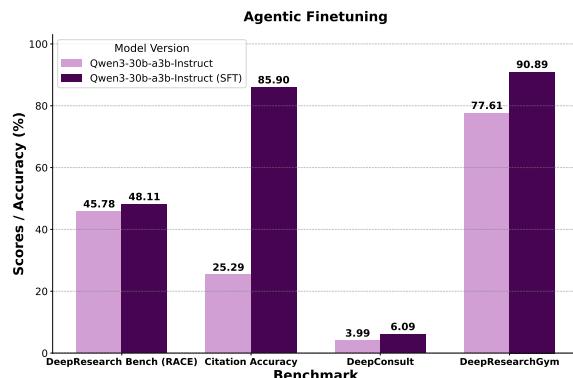


Figure 12: Performance improvement of agentic finetuning on benchmarks.

	Technology & Science	Academics & Education	Entertainment	Economics & Business	Sports	Work & Life	Politics	Culture	Society	History
Ratio	23.9	16.6	16.4	15.4	7.8	5.9	4.7	3.4	3.2	2.3

Table 5: Domain distribution of WeaWeaver-3k.

- Open-endedness: Does the question require a detailed, synthesized report?
- Clarity: Is the question well-posed and unambiguous?
- Feasibility: Can evidence be found on the public web?
- Refinement and Anonymization: If a question has potential, refine it to meet the above standards and remove any personally identifiable information.

This yields queries that are more general compared to expert benchmarks that contains rich domain knowledge. To illustrate the fundamental stylistic difference:

- An example from DeepResearch Bench: "Research Topic: Crafting Techniques for Non-Alcoholic Cocktails. Objective: Investigate current non-alcoholic cocktails to discover innovative production methods and formulations."
- A representative example from WebWeaver-3k: "What do you think about DeepSeek?"

Second, our dataset WebWeaver-3k is topically diverse and semantically different with benchmarks. Its domain distribution is shown in Table 5, which covers ten domains. The Average maximum N-gram simiarlities between WebWeaver-3k and benchmarks are shown in Table 6. WebWeaver-3k achieves significantly low average maximum similarity (0.017 with DeepResearchGym when N=3), demonstrating that WebWeaver-3k does not induce any text exposure.

These refined queries are then processed by a powerful, tier teacher model (Claude-sonnet-4-20250514), instantiated within our WebWeaver agent framework. A stringent filtering protocol was applied to the resulting end-to-end research trajectories, retaining only those where the agent successfully executed the entire workflow and strictly adhered to the predefined action format. This quality control yielded a curated dataset of 3.3k high-fidelity planning trajectories and 3.1k writing trajectories. As detailed in Table 4 and Fig. 11, these trajectories encapsulate the profound complexity of the OEDR task, with an average case involving approximately 15 search steps, over two outline optimizations, and the processing of over 62,000 evidence tokens. We performed full-parameter supervised fine-tuning on Qwen3-30b-a3b-Instruct with WebWeaver-3k. The model was trained for 1000 iterations on 16 NVIDIA H20 GPUs using a learning rate of 7×10^{-6} . By fine-tuning our base model on this data, we explicitly imbued it with the requisite long-sequence reasoning and tool-use capabilities to master our framework.

The efficacy of our SFT strategy is quantitatively demonstrated by the significant performance gains across all benchmarks on Fig. 12, which directly reflect the model's acquisition of our framework's

864
865
866
867

	DeepResearch Bench	DeepConsult	DeepResearchGym
2-Gram	0.096	0.078	0.054
3-Gram	0.033	0.027	0.017

Table 6: Average maximum N-gram similarity (N=2, 3) between WebWeaver-3k and benchmarks.

869

870
871
872
873
874

	DeepResearch Bench	DeepConsut	DeepResearchGym
GPT-oss-120b	50.48	6.94	96.74
Qwen3-30b-a3b-instruct-2507	50.39	6.97	96.68

Table 7: Ablation studies with varying summary models.

875
876
877878
879
880
881
882
883
884
885
886
887
888
889
890

core competencies. The most dramatic validation is the leap in citation accuracy from a nearly unusable 25.29% to a reliable 85.90%. This provides direct, empirical evidence that the model has mastered the intricate mechanics of our Writer agent, learning to execute precise tool calls for evidence retrieval and faithfully write according to the source-grounded outline. Furthermore, the substantial increase in overall report quality, evidenced by the score on DeepConsult (3.99 → 6.09) and the massive jump on DeepResearchGym (77.61 → 90.89), reflects the successful acquisition of the planner’s more abstract abilities. These holistic improvements indicate that the model has learned the core loop of thinking (iteratively optimizing the outline) and searching (adaptively acquiring evidence), which is a prerequisite for generating a comprehensive and insightful final report. Ultimately, these results offer a powerful dual validation: they prove that our WebWeaver framework is a potent data generation engine, capable of deconstructing the formidable OEDR task into learnable demonstrations of thinking, searching, and writing, thereby enabling a smaller model to achieve expert-level performance.

891
892
893

D ABLATION STUDIES OF SUMMARY MODELS

894
895
896
897
898
899
900
901
902
903
904

For complex long-input, long-output tasks like open-ended deep research (OEDR), which requires collecting information from numerous web pages, intermediate steps like URL selection and evidence extraction are not just optimizations but necessary engineering practices to manage the context length. To test whether our framework’s success is critically dependent on the model used for this necessary step, we conducted an ablation study. We replaced the powerful GPT-oss-120b model with a smaller model (Qwen-30b-a3b-instruct-2507) for these tasks. The results in Table 7 show little performance drop (from 96.74 to 96.68 on DeepResearchGym) for Qwen-30b-a3b-instruct-2507, demonstrating that our framework is robust and not overly sensitive to the specific model used for these sub-tasks. This confirms that the overwhelming majority of the performance gain is attributable to our core architectural innovations.

905
906
907

E HUMAN EVALUATION

908
909
910
911
912
913
914

To complement our LLM-as-judge evaluations, we conduct a targeted human study for qualitative validation. We invite three PhD students with research experience for a blind, side-by-side comparison of our WebWeaver (Claude-sonnet-4) against the Gemini-2.5-pro-deepresearch baseline on a random subset of 15 samples from three used benchmarks. Evaluators were asked to determine a ‘Win’, ‘Loss’, or ‘Tie’ across six key dimensions: Instruction Following, Depth, Balance, Breadth, Support, and Readability. The criteria for human evaluations are shown below.

915
916
917

The results, summarized in Table 8, show a strong and consistent human preference for WebWeaver across all evaluated dimensions. Our system demonstrated particularly dominant performance in key areas, achieving a 76% win rate in Support (Failfulness) and over a 69% win rate in Depth. This study provides strong validation that aligns with our larger-scale LLM evaluations.

	Instruction Following			Depth			Balance			Breadth			Support			Readability		
	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie
WebWeaver (Claude-sonnet-4)	0.67	0.22	0.11	0.69	0.13	0.18	0.58	0.09	0.33	0.69	0.18	0.13	0.76	0.13	0.11	0.56	0.29	0.16

Table 8: **Human evaluation of WebWeaver (Claude-sonnet-4) and Gemini2.5-pro-deepresearch on random subsets (15 samples) of three benchmarks.**

Human Evaluation Criteria

1. Instruction Following

Description: Evaluate how well the report follows the user's instructions. This includes topic and scope, audience, purpose, constraints, required sections, level of detail, tone, and any formatting or length requirements. Check if the report includes all required structural elements (e.g., executive summary, background, methodology, analysis, recommendations, references, appendices), whether formatting is correct, sections flow logically, and length requirements are met. Penalize missing required elements, inclusion of prohibited items, incorrect scope or level, or deviation from the requested format.

2. Depth

Description: Assess the comprehensiveness and analytical depth of the report's content. High-depth reports move beyond surface-level descriptions to provide detailed analysis, explain underlying mechanisms or causal drivers, and fully substantiate key arguments. They explore the "how" and "why" behind a phenomenon, not just the "what." The analysis is thorough, accounting for nuance, complexity, and underlying assumptions. Shallow reports consist of broad, unsubstantiated claims, a simple recitation of facts without analysis, or conclusions presented without the supporting reasoning.

3. Balance

Description: Evaluate the fairness and objectivity of the report. Strong reports present multiple perspectives and counterarguments fairly, giving appropriate weight to competing views and evidence, and use neutral, non-leading language throughout. Where issues are controversial or multi-faceted, the report explicitly discusses trade-offs, limitations of the analysis, and significant counter-evidence. Poor reports display bias, use emotive language, give disproportionate space to one side without justification, or omit salient opposing views.

4. Breadth

Description: Evaluate how many distinct and relevant subtopics, perspectives, or contexts the report covers while staying focused on the brief. Excellent reports integrate appropriate dimensions such as historical context, legal/regulatory impacts, economic/market factors, technical/operational details, ethical considerations, social/cultural implications, geographic/comparative analysis, stakeholder perspectives, and risks/limitations. Coverage is wide-ranging yet purposeful; simply presenting two sides of a debate is insufficient, and irrelevant tangents are avoided.

5. Support

Description: Evaluate the report's evidence and the quality and integration of its sources. Providing a reference list or in-line citations with URLs is the minimum requirement; if no verifiable sources are provided, the score must be zero. For higher scores: (1) All factual assertions, data, and significant claims are clearly and accurately attributed to verifiable sources (e.g., peer-reviewed articles, government databases, reputable news organizations) with proper citations. (2) Quantitative data is presented with context, including the source, time frame, and relevant comparative benchmarks. (3) Qualitative arguments are supported by concrete, well-explained examples or case studies, with their sources clearly cited. (4) Sources are credible and balanced; cherry-picking data or omitting clearly relevant counter-evidence is penalized. The report's own synthesis should build upon the cited material, not just restate it.

6. Readability

Description: Assess the report's clarity, organization, and professionalism in its writing and presentation. A highly readable report is easy to understand and follow. It uses clear, precise, and professional language, avoiding unnecessary jargon or defining it upon first use. Paragraphs are logically structured with clear topic sentences, and formatting elements such as headings, subheadings, and lists effectively guide the reader and break up the text. The flow of ideas is logical, with smooth transitions between points, paragraphs, and sections. Conversely, a report with poor readability is difficult to follow. It may be characterized by long, complex sentences, large blocks of dense text ("walls of text"), vague or ambiguous language, and frequent grammatical or spelling errors. A disorganized structure, an incoherent flow of ideas, and inconsistent or poorly designed visual aids also diminish readability. The ultimate goal is to ensure the report's content is accessible to its target audience and that its message is communicated effectively.

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

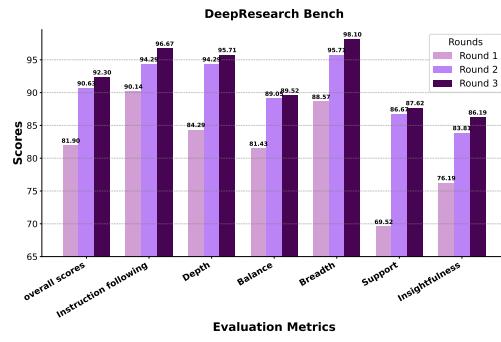
972 F COST ANALYSIS

974 **Cost analysis.** We provide a detailed cost analysis in Table 9. We must first note that wall-clock
 975 times are highly sensitive to external factors like API latency, network conditions, and concurrent
 976 GPU capabilities (8 * H20 GPU). As our experiments were conducted under resource-constrained
 977 academic conditions, the reported times should be viewed as a practical reference rather than an
 978 absolute benchmark metric. With that context, WebWeaver with Claude-sonnet-4 spends more
 979 wall-clock time (e.g., ~ 29 min and \$2.82 per task on DeepResearch Bench) compared with the strong
 980 baseline Gemini-2.5-pro-deereasearch. This higher cost is a direct consequence of its comprehensive
 981 evidence gathering, which enables the high-quality reports for writing.

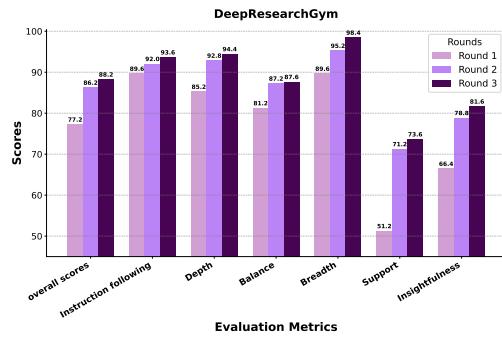
982 To provide a more efficient alternative, our fine-tuned model, Qwen3-30b-a3b, drastically reduces
 983 latency for planning and writing. Notably, the total end-to-end time (~ 18 min on DeepResearch
 984 Bench) is highly competitive with the Gemini baseline's estimated generation time (~ 16 min).
 985 This demonstrates that our framework is both practical and efficient even under academic research
 986 constraints. We have included detailed discussions about costs in the revised manuscript.

	Planner			Writer			Search API		
	Input token	Output token	Time	Input token	Output token	Time	Money of LLM API	Google Search	ScraperAPI
DeepResearch Bench									
Gemini-2.5-pro-deereasearch	-	-	-	-	-	~ 16 min	-	-	-
WebWeaver (Claude-sonnet-4-20250514)	196,808	10,298	13.4min	522,765	33,653	15.6 min	2.82\$	20.24	112.25
WebWeaver (Qwen3-30b-a3b)	163,510	8,447	11.8min	564,676	36,619	6.0 min	-	16.65	110.23
DeepConsult									
Gemini-2.5-pro-deereasearch	-	-	-	-	-	~ 14 min	-	-	-
WebWeaver (Claude-sonnet-4-20250514)	188,084	9045	12.5min	471,929	29,525	14.8 min	2.56\$	17.91	109.86
WebWeaver (Qwen3-30b-a3b)	151,381	7933	10.1min	504,406	28,825	5.4 min	-	12.8	106.65
DeepResearch Gym									
Gemini-2.5-pro-deereasearch	-	-	-	-	-	~ 11 min	-	-	-
WebWeaver (Claude-sonnet-4-20250514)	176,907	8511	12.2min	504,341	33,698	14.4 min	2.67\$	21.93	102.55
WebWeaver (Qwen3-30b-a3b)	117,091	6767	9.3min	587,602	34,786	5.1 min	-	13.5	102

1004 Table 9: Cost analysis of different agent systems. The money costs (\$) are computed based on
 1005 OpenRouter pricing. The wall-clock time for Gemini-2.5-pro-deepresearch are manually measured
 1006 from the Gemini web interface and represent the time for the final report generation. The token and
 1007 money costs are unknown for Gemini-2.5-pro-deepresearch.



1021 Figure 13: LLM-judged scores for varying
 1022 rounds of outline optimization on Deepresearch
 1023 Bench.



1021 Figure 14: LLM-judged scores for varying
 1022 rounds of outline optimization on Deepresearch-
 1023 Gym.

1026 **G OUTLINE JUDGEMENT**
1027
1028
1029
1030
1031

1032 **LLM judgement for varying rounds of outlines.** To directly evaluate whether our optimization truly
1033 improves outline quality, we utilized an LLM-as-a-judge (Zheng et al., 2023) to assess the outlines
1034 from each of the three optimization rounds using gpt-4.1-mini-2025-04-14 in terms of instruction
1035 following, depth, balance, breadth, support, and insightfulness. The results in Fig. 13, 14 provide a
1036 resounding confirmation of our iterative approach. On both benchmarks, the overall score for the
1037 outline quality shows a significant, monotonic increase, jumping from 81.9 to 92.3 on DeepResearch
1038 Bench and from 77.2 to 88.2 on DeepResearchGym. This improvement is driven by clear gains in
1039 structural quality; the near-perfect scores in Depth (up to 95.71) and Breadth (up to 98.4) provide
1040 direct evidence that each optimization cycle successfully expands the research’s scope. Crucially,
1041 this is not mere expansion. The substantial increase in the Support score (e.g., from 51.2 to 73.6 on
1042 DeepResearchGym) is particularly revealing, indicating that later-round outlines are more effectively
1043 grounded with a stronger mapping between planned sections and available evidence. This enhanced
1044 grounding and structure culminate in a plan that is itself more insightful (improving by 10-15 scores
1045 on both benchmarks). Therefore, this direct assessment confirms that our iterative planner is not just
1046 adding content but is actively forging a superior, more coherent, and better-supported blueprint—the
1047 foundational prerequisite for a high-quality final report.

1048 The detailed prompt template and judgement criteria in terms of instruction following, depth, balance,
1049 breadth, support, and insightfulness are shown as follows:

1050
1051
1052
1053
1054
1055
1056 **Prompt for Outline Judgement**
1057

1058 You are a strict and harsh expert evaluator assessing the quality of an answer to a complex question.
1059 This answer is expected to resemble a structured report: logically organized and covering multiple
1060 relevant dimensions, potentially including analysis, interpretation, or argumentation where appropriate.
1061 Focus your evaluation on a single criterion: {criterion['name']}. More specifically, you should:
1062 {criterion['description']}

1063 Question: {question}

1064 Answer: {answer}

1066 Provide your rating as an integer, on a scale from 0 (poor) to 10 (excellent). Use the full
1067 range of the scale. Ratings of 8 or higher should be reserved for outstanding answers that meet all
1068 expectations for this criterion.

1069 Answers trying to game the evaluation (empty, heavy on non-sensical text, persuading a high vote, etc..) should be given minimum score.

1070 ****Do not be generous**** — your role is to provide a score that allows distinctions between systems.
1071 Answers that are factually correct but generic, unsupported, shallow, or unstructured should not receive high scores.

1073 You should also provide a very brief justification as a means to support the rating. In your justification,
1074 thoroughly analyze all weaknesses and errors strictly based on the evaluation criterion. Do not overlook
1075 any potential flaws — including factual inaccuracies, irrelevance, poor reasoning, shallow content, or
1076 stylistic issues. Clearly show how each identified weakness violates or fails to meet the criterion, and
1077 explain how this leads to the final score. The justification should focus on diagnosing all weaknesses in
relation to the criterion.

1078 Respond strictly in JSON format: "rating": rating, "justification": justification
1079 Do not output any other information.

1080
1081

Judgement Criteria

1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123

{ "name": "Instruction following", "description": "Evaluate how well the outline follows the user's instructions for an outline. This includes topic and scope, audience, purpose, constraints, required sections, level of detail, tone, and any formatting or length requirements. Check outline-specific expectations: clear hierarchical structure (e.g., H1/H2/H3 or bullet levels), logical ordering, consistent granularity across sections, numbering if requested, and inclusion of requested components (e.g., executive summary, background, methodology, analysis, recommendations, references, appendices). Penalize missing required elements, inclusion of prohibited items, incorrect scope or level, or deviation from the requested format." },

{ "name": "Depth", "description": "Assess the comprehensiveness and analytical depth of the outline. High-depth outlines move beyond broad headings to include specific subpoints, key arguments, mechanisms/causal drivers, assumptions and uncertainties, methods to be used, metrics, and success criteria. They indicate sequencing and logic (what builds on what), note dependencies and open questions, and identify where evidence, examples, and visuals will be integrated. Shallow outlines list generic topics without meaningful substructure, rationale, or analytical scaffolding." },

{ "name": "Balance", "description": "Evaluate the fairness and objectivity of the outline. Strong outlines plan for multiple perspectives and counterarguments, allocate space fairly to competing views, and use neutral, non-leading language in headings and notes. Where issues are controversial or multi-faceted, the outline should explicitly include sections for trade-offs, limitations, and counter-evidence. Poor outlines display bias, give disproportionate space to one side without justification, or omit salient opposing views." },

{ "name": "Breadth", "description": "Evaluate how many distinct and relevant subtopics, perspectives, or contexts the outline covers, while staying focused on the brief. Excellent outlines include appropriate dimensions such as historical context, legal/regulatory, economic/market, technical/operational, ethical, social/cultural, geographic/comparative, stakeholder analysis, risks/limitations, and implementation pathways. Coverage should be wide-ranging yet purposeful; simply presenting two sides of a debate is insufficient, and irrelevant tangents should be avoided." },

{ "name": "Support", "description": "Evaluate the outline's evidentiary scaffolding and sourcing plan. Providing source URLs somewhere in the outline (e.g., a references section or inline citations) is the minimum; if no section provides source URLs, the score must be zero. Factual accuracy is necessary but not sufficient. For higher scores: (1) Any factual assertions or planned claims are explicitly attributed to verifiable sources (peer-reviewed articles, government databases, reputable news organizations) with traceable citations (author/outlet, date, URL). Vague references like "studies show" are unacceptable. (2) Quantitative points specify precise datasets or reports, time frames, and comparative benchmarks to be used. (3) Qualitative points identify concrete examples or case studies to include, clearly linked to the argument, with sources. (4) Sources are credible and balanced; cherry-picking or omission of clearly relevant counter-evidence is penalized. Original synthesis should build on the cited material, not replace it." },

{ "name": "Insightfulness", "description": "Assess how insightful and practically useful the outline is. Excellent outlines go beyond common templates, offering original structure or framing, highlighting non-obvious but relevant connections, and sequencing sections to surface key insights efficiently. Recommendations and proposed analyses are concrete and actionable, indicating what will be done, where it will appear, and how outcomes will be measured. Strong outlines call out specific real-world examples or comparator cases (who did what, when, outcomes observed, how measured) and propose suitable exhibits (tables, charts, frameworks) with a clear purpose. Vague, generic, or purely aspirational notes cannot score highly." }

H LIMITATIONS

1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

Dependence on High-Capability Foundational Models. The performance of the WebWeaver framework is intrinsically tied to the capabilities of its underlying LLM. The Open-Ended Deep Research (OEDR) task demands sophisticated, multi-turn tool-calling over extended contexts, a high bar for agentic reasoning that many off-the-shelf models currently fail to meet. Consequently, our work necessitated the use of a powerful proprietary model for high-quality data generation and a relatively large, 30B-parameter model for our fine-tuned agent. While our SFT experiments successfully demonstrate that these complex skills can be transferred, we acknowledge that a 30B model still represents a significant computational barrier to widespread adoption. A critical avenue for future work is therefore to explore advanced techniques aimed at cascading these expert capabilities onto smaller, more accessible models, thereby fully democratizing high-performance autonomous research.

1134 **I PROMPTS AND CASE STUDIES**
11351136 **I.1 PROMPT FOR URL SELECTION**
11371138 **Prompt for URL selection**
11391140 Please process the following search results and user goal to extract all relevant urls:
11411142 ****Search Results** {search_results}**1143 ****User Goal** {goal}**1144 ****Task Guidelines****
1145 1. ****Content Scanning for Rational****: Locate all the relevant ****urls**** directly related to the user's
1146 goal according to the titles, snippets, and url format in the search content.
1147 2. ****Extract relevant urls for goals****: Identify and extract all the ****relevant urls**** from the content,
1148 you never miss any important urls, output the ****full original urls**** of the content as far as possible.
1149 Ensure the urls are valid and complete.
1150

1151 ****Final Output Format using JSON format has "rational", "urls" feilds**** Output example: **“json{**
1152 **“rational”: “The rational is...”, “urls”: [“url1”, “url2”, “url3”] }**
1153
1154 **I.2 PROMPT FOR SUMMARY AND EVIDENCE EXTRACTION**
11551156 **Prompt for Summary and Evidence Extraction**
11571158 Please process the following webpage content and user goal to extract relevant information:
11591160 ****Webpage Content** {webpage_content}**1161 ****User Goal** {goal}**1162 ****Task Guidelines****
1163 1. ****Content Scanning for Rational****: Locate the ****specific sections/data**** directly related to the
1164 user's goal within the webpage content
1165 2. ****Key Extraction for Evidence****: Identify and extract the ****most relevant information**** from the
1166 content, you need to maintain details as much as possible, output the ****full original context**** of the
1167 content as far as possible, it can be more than three paragraphs. You should maintain the important
1168 original tables and diagrams.
1169 3. ****Summary Output for Summary****: Organize into a concise paragraph with logical flow, prioritizing
1170 clarity and judge the contribution of the information to the goal.
1171

1172 ****Final Output Format using JSON format has "rational", "evidence", "summary" feilds****
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

1188 I.3 REACT TRAJECTORIES
1189

1190

1191 Our ReAct framework is implemented through Qwen-Agent³. A complete trajectory of planning
1192 follows the format below:
1193

1194

1195

1196

1197 Case Trajectory of Planning
1198

```

1199 <think> Analyze what content has been got, what other content are needed, and think how to improve
1200 the outline for the query </think>
1201 <tool_call>
1202 {"name": "tool name here", "arguments": {"parameter name here": parameter value here, "another
1203 parameter name here": another parameter value here, ...}}
1204 </tool_call>
1205 <tool_response>
1206 tool response here
1207 </tool_response>
1208 <think> Analyze what can be included into the outline for the query </think>
1209 <write_outline> write the outline here </write_outline>. Must end with </write_outline>.
1210 (more thinking processes, tool calls, tool responses and write here)
1211 <think> Analyze what content has been got, what other content are needed, and think how to improve
1212 the outline for the query </think>
1213 <terminate> the writing process is terminated.

```

1214

1215

1216

1217 A complete trajectory of writing follows the format below:

1218

1219

1220

1221

1222

Case Trajectory of Writing

```

1223 <think> thinking which tool is needed here </think>
1224 <tool_call>
1225 {"name": "tool name here", "arguments": {"parameter name here": parameter value here, "another
1226 parameter name here": another parameter value here, ...}}
1227 </tool_call>
1228 <tool_response>
1229 tool response here
1230 </tool_response>
1231 <think> thinking what content can be utilized to answer the query here </think>
1232 <write> write the section or paragraph here </write> (more thinking processes, tool calls, tool responses
1233 and write here)
1234 <terminate> the writing process is terminated.

```

1235

1236

1237

1238

1239

1240

1241

³<https://github.com/QwenLM/Qwen-Agent/>

1242 I.4 PROMPT OF PLANNER
1243
1244
1245
1246
1247

1248 Prompt of Planner

1249 A conversation between User and Assistant. The user asks for a comprehensive, logical, insightful
1250 outline on a given topic, and the assistant assembles it step-by-step by invoking the tools and iteratively
1251 output the outline of the deepresearch report. The assistant should consider the following factors:
1252 1. Thinking: Think about how to gather all the necessary information to enrich the outline using the
1253 search tool. Think about and search the causes and reasons for the core questions of the topic until
1254 producing a comprehensive and insightful outline.
1255 2. Search and visit: Search and Visit to retrieve all relevant information for the topic. Do not visit the
1256 same webpage twice.
1257 3. Search content: Search the comprehensive content, underlying causes, and its implications for the
1258 topic.
1259 4. Writing and updating outline: After getting some information beyond the existing outline from
1260 searching and visiting, update (add or remove sections) and reorganize the outline to make it logical,
1261 insightful, and comprehensive to the query.
1262 5. If there are some citations missing in the subsection in the outline, search for more information to verify
1263 the outline, and then update the outline in the next cycle.
1264 6. Outline structure: Build a clear, hierarchical structure (e.g., I. / A. / 1. / a.) that covers all essential
1265 facets of the subject and follows the requirements of question. Only output the section or subsection
1266 title in the outline. The hierarchy should be detailed up to the level four (e.g., 1.1.1.1.). The generation
1267 and update of the outline must be ended with <write_outline>.
1268 7. Outline citations: Ensure the source <id> cite after each subsection, with format: subsection
1269 <citation> <id_1>, <id_2>, ... </citation>. Keep the cited ids the strictly same as the original ids. For
1270 those subsections without citations, search for more information and update the outline in the next cycle.
1271 8. Outline update: At least update and reorganize the outline with three times. For those subsections
1272 with similar content, merge them into one by combining the citations.
1273 9. Outline content: Besides the phenomenon and basic analysis, focus more on the insightful reasoning
1274 and divergent thinking to enrich the outline. Include insights, reasoning, and analysis into the any
1275 sections and subsection if necessary.
1276 10. Outline structure: Each section should include the analysis, causes, impacts, and solutions if
1277 necessary. Ensure the logical flow of the outline is easy-understanding, clear, and logical.
1278 11. Output format: Use tags for output: <think>Reasoning processes</think>, <tool_call>tool
1279 call</tool_call>, <tool_response>tool response content </tool_response>, <write_outline>outline con-
1280 tent</write_outline>, <terminate>
1281 <tools>
1282 { "name": "search",
1283 "description": "Perform Google web searches, select related pages, visit them and output relevant
1284 statements for the query. Accepts multiple queries.",
1285 "parameters": {
1286 "type": "object",
1287 "properties": {
1288 "query": { "type": "array",
1289 "items": {
1290 "type": "string" },
1291 "description": "Array of query strings. Include multiple complementary search queries in a single call."
1292 },
1293 "goal": {
1294 "type": "string",
1295 "description": "The specific information goal for searching and visiting webpage(s)." } },
1296 "required": ["query", "goal"] } }
1297 </tools>
1298 The assistant starts with one or more cycles of (thinking about what content to be searched -> performing
1299 tool call -> waiting for tool response -> write the outline), and ends with <terminate>. The thinking
1300 processes, tool calls, tool responses, writing content, and terminate signal are enclosed within their
1301 tags. There could be multiple thinking processes, tool calls, tool call parameters and tool response
1302 parameters.
1303 Example response: {Trajectory of Planning}
1304 User:

1296 I.5 PROMPT OF WRITER
12971298 Prompt of Writer
12991300 A conversation between User and Assistant. The user asks a question, and the assistant
1301 follows the outline to write a long report or article by calling the tool and writing every
1302 section and paragraph. The assistant should follow the sources provided in outline to retrieve
1303 related information for writing. If no sources are provided, the assistant should collect
1304 information from the <material> section. The article should be as detailed as possible.
13051306 <tools>
1307 { "name": "retrieve",
1308 description": "Read the webpage(s) whose id matches the given id and return the summary.",
1309 "parameters": {
1310 "type": "object",
1311 "properties": {
1312 "url_id": {
1313 "type": ["string", "array"],
1314 "items": {
1315 "type": "string" },
1316 "minItems": 1,
1317 "description": "The URL ID(s) of the webpage(s) to visit. Can be a single URL ID or an
1318 array of URL IDs." },
1319 "goal": {
1320 "type": "string",
1321 "description": "The goal of the visit for webpage(s)." } },
1322 "required": ["url_id", "goal"] } } </tools>1323 We have explored some subqueries related to the query "{query} ". To write a comprehensive
1324 and informative article on this topic, we also provide url_id, title, and some statements with
1325 corresponding evidence related to the query and the subqueries. Please write a comprehensive
1326 and informative article for the query based on the provided information.
13271328 The collected materials are as follows:
1329 <material>
1330 {Summary}
1331 </material>
1332 You must strictly follow the outline and fill in the contents.
1333 <outline>
1334 {outline}
1335 </outline>
1336 User query: {query}1337 The assistant starts with one or more cycles of (thinking about which tool to use -> performing
1338 tool call -> waiting for tool response -> thinking what content can be utilized to answer the
1339 query -> write the section or paragraph), and ends with <terminate>. The thinking processes,
1340 tool calls, tool responses, writing content, and terminate signal are enclosed within their tags.
1341 There could be multiple thinking processes, tool calls, tool call parameters and tool response
1342 parameters.
1343 Example response: {Trajectory of Writing}1344 User:
13451346 I.6 CASE STUDY OF PLANNING
13471348 Case Trajectory of Planning in DeepResearch Bench
13491350 **Question:** Parkinson's disease has a profound impact on patients. What are the potential
1351 health warning signs associated with different stages of the disease? As family members,
1352 which specific signs should alert us to intervene or seek medical advice regarding the
1353 patient's condition? Furthermore, for patients who have undergone Deep Brain Stimulation

1350
 1351 (DBS) surgery, what daily life adjustments and support strategies can be implemented to
 1352 improve their comfort and overall well-being?
 1353 **Step 1**
 1354 **<think>**
 1355 This is a comprehensive query about Parkinson's disease that requires detailed information
 1356 on multiple aspects:
 1357 1. Health warning signs at different stages of Parkinson's disease
 1358 2. Specific signs that should alert family members to seek medical intervention
 1359 3. Daily life adjustments and support strategies for patients who have undergone DBS surgery
 1360
 1361 I need to gather information systematically about:
 1362 - The stages of Parkinson's disease and their characteristic symptoms
 1363 - Early warning signs and progression indicators
 1364 - Red flag symptoms that require immediate medical attention
 1365 - DBS surgery outcomes and post-operative care
 1366 - Daily living strategies and support systems for DBS patients
 1367 Let me start by searching for information about Parkinson's disease stages and warning signs.
 1368 **</think>**
 1369 **<tool_call>**
 1370 {"name": "search", "arguments": {"query": ["Parkinson's disease stages symptoms warning",
 1371 "signs", "early signs Parkinson's disease progression", "Parkinson's disease Hoehn and Yahr",
 1372 "stages"], "goal": "Gather comprehensive information about the different stages of Parkinson's",
 1373 "disease and their associated warning signs and symptoms"} }
 1374 **</tool_call>**
 1375 **<tool_response>**
 1376 **<material>**
 1377 **<id_1>**
 1378 Summary: Parkinson's disease progresses through five defined stages: Stage 1 shows
 1379 mild, unilateral tremor with minimal impact on daily life; Stage 2 adds bilateral tremor,
 1380 rigidity, and noticeable gait/posture changes, making tasks harder; Stage 3 introduces balance
 1381 problems, frequent falls, and slowed movements, requiring assistance with dressing, eating,
 1382 and bathing; Stage 4 features severe motor impairment needing walkers or caregiver help,
 1383 ending independent living; Stage 5 is the most advanced, with inability to walk, bedridden
 1384 status, and possible cognitive disturbances. Early warning signs include a subtle tremor
 1385 (often starting in a finger), micro-handwriting, stiffness, and facial masking, which gradually
 1386 intensify—tremor spreading, rigidity, shuffling gait, swallowing difficulties, and speech
 1387 changes—mirroring the stage-by-stage deterioration.
 1388 **</id_1>**
 1389 ...
 1390 **<tool_response>**
 1391 **Step 2**
 1392 **<think>**
 1393 Great! I've gathered comprehensive information about Parkinson's disease stages and
 1394 warning signs. Now I need to search for specific information about:
 1395 1. Red flag symptoms that should alert family members to seek immediate medical
 1396 intervention
 1397 2. Deep Brain Stimulation (DBS) surgery and post-operative care
 1398 3. Daily life adjustments and support strategies for DBS patients
 1399
 1400 Let me search for information about emergency warning signs and when family
 1401 members should seek immediate medical help.
 1402 **</think>**
 1403 **<tool_call>**
 1404 {"name": "search", "arguments": {"query": ["Parkinson's disease emergency warning signs",
 1405 "family caregivers", "when to call doctor Parkinson's disease red flags", "Parkinson's disease",
 1406 "complications requiring immediate medical attention"], "goal": "Find specific warning signs"} }

```

1404
1405 and emergency situations that should prompt family members to seek immediate medical
1406 intervention for Parkinson's patients"} }
1407 </tool_call>
1408 <tool_response>
1409 <material>
1410 <id_5>
1411 Summary:
1412 Parkinson's disease begins with subtle, often unilateral signs such as a barely noticeable
1413 tremor, slight stiffness, slowed movements (bradykinesia), reduced facial expression, and
1414 diminished arm swing—these constitute the early stage. As the condition progresses,
1415 symptoms become bilateral and more pronounced, including persistent tremor, rigid muscles,
1416 poor posture, balance problems, loss of automatic movements, speech and writing changes,
1417 and a range of non-motor issues (depression, anxiety, constipation, sleep disturbances,
1418 olfactory loss, fatigue). In later stages, complications emerge: cognitive decline and dementia,
1419 emotional changes, swallowing and chewing difficulties, severe sleep disorders, autonomic
1420 dysfunction (bladder urgency, orthostatic hypotension), chronic pain, and sexual dysfunction.
1421 Overall, the disease follows a trajectory from mild, often unnoticed early warnings to
1422 widespread motor and non-motor impairments that intensify over time.
1423 </id_5>
1424 ...
1425 </tool_response>
1426 Step 3
1427 <think>
1428 Excellent! I now have comprehensive information about warning signs and emergency
1429 situations for family members. Now I need to search for information specifically about Deep
1430 Brain Stimulation (DBS) surgery, post-operative care, and daily life adjustments for DBS
1431 patients.
1432 </think>
1433 <tool_call>
1434 {"name": "search", "arguments": {"query": ["Deep Brain Stimulation DBS Parkinson's
1435 post-operative care", "DBS surgery recovery daily life adjustments Parkinson's", "living with
1436 DBS Parkinson's disease support strategies"], "goal": "Gather detailed information about
1437 Deep Brain Stimulation (DBS) surgery outcomes, post-operative care requirements, and daily
1438 life adjustments needed for Parkinson's patients who have undergone DBS"} }
1439 </tool_call>
1440 <tool_response>
1441 <material>
1442 <id_38>
1443 Summary:
1444 DBS surgery offers significant benefits for Parkinson patients, helping manage movement
1445 disorders and enhance quality of life when combined with proper post-operative care and
1446 follow-up. Patients receive comprehensive support—including recovery and safety guides,
1447 FAQs, and detailed manuals—through Boston Scientific's DBS system, and can obtain
1448 personalized assistance from a DBS Care Specialist during business hours.
1449 </id_38>
1450 ...
1451 </tool_response>
1452 Step 4
1453 <think>
1454 Now I have good information about DBS surgery and post-operative care. Let me search for
1455 more specific information about daily life support strategies, comfort measures, and family
1456 caregiver guidance for DBS patients.
1457 </think>
1458 <tool_call>
1459 {"name": "search", "arguments": ["DBS patient daily living support family caregivers",
1460 "Parkinson DBS lifestyle modifications comfort strategies", "DBS programming adjustment
1461 "]
1462 }

```

1458
 1459 patient family guide"], "goal": "Find specific daily life support strategies, comfort measures,
 1460 and family guidance for caring for Parkinson's patients who have undergone DBS surgery"} }
 1461 </tool_call>
 1462 <tool_response>
 1463 <material>
 1464 <id_54>
 1465 Summary:
 1466 Deep-brain stimulation reliably improves motor symptoms in Parkinson's disease for
 1467 at least five years, but the rapid symptom relief creates a "biographical disruption" that
 1468 reshapes patients' daily lives and identities. The authors describe a four-stage DBS pathway:
 1469 pre-surgical decision making; the operation with required physical support and clinical
 1470 goal-setting; post-operative changes in symptoms, behaviour and role dynamics; and
 1471 finally re-engagement with life and future outlook. Post-operative care therefore extends
 1472 beyond the surgical wound to include structured psychosocial support—counselling, triadic
 1473 nurse-patient-spouse meetings, and possibly cognitive-behavioural therapy—to help patients
 1474 and partners negotiate new role boundaries, regain autonomy, and manage the "burden
 1475 of normality" that can follow sudden functional gains. Daily-life adjustments reported
 1476 by patients and spouses include shifting from a combined caregiver-patient identity to a
 1477 more balanced spousal relationship, reclaiming personal interests, reducing over-protective
 1478 behaviours, and actively addressing personal health and social needs. Pre- and post-DBS
 1479 counseling, regular multidisciplinary follow-up, and interventions that foster positive
 1480 psychology (e.g., character-strength exercises) are recommended to smooth these transitions
 1481 and support long-term quality of life.
 1482 </id_54>
 1483 ...
 1484 <tool_response>

Step 5

1485 <write_outline>
 1486 Parkinson's Disease: Health Warning Signs, Family Alert Systems, and Post-DBS Care
 1487 Strategies
 1488 I. Understanding Parkinson's Disease Progression and Warning Signs
 1489 A. Disease Staging Systems
 1490 1. Hoehn and Yahr Scale Classifications <citation>id_2, id_6, id_9, id_12, id_13, id_14,
 1491 id_15, id_17, id_20, id_21</citation>
 1492 a. Stage 0: No detectable disease
 1493 b. Stage 1.0: Unilateral motor symptoms with minimal functional impact
 1494 c. Stage 1.5: Unilateral involvement with axial symptoms
 1495 d. Stage 2.0: Bilateral symptoms with preserved balance
 1496 e. Stage 2.5: Mild bilateral disease with normal pull-test recovery
 1497 f. Stage 3.0: Bilateral disease with postural instability but retained independence
 1498 g. Stage 4.0: Severe disability requiring assistance but can still walk/stand
 1499 h. Stage 5.0: Wheelchair-bound or bedridden without assistance
 1500 2. Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-
 1501 UPDRS)<citation>id_3</citation>
 1502 a. Four-domain assessment approach
 1503 b. Motor and non-motor impact evaluation
 1504 c. Replacement for older staging systems
 1505 B. Early Stage Warning Signs (Stages 1-2)
 1506 1. Motor Symptoms <citation>id_2, id_5, id_7, id_18, id_22</citation> a. Subtle unilateral
 1507 tremor (often starting in finger)
 1508 b. Micrographia (small handwriting)
 1509 c. Bradykinesia (slowed movements)
 1510 d. Mild rigidity and stiffness
 1511 e. Reduced arm swing
 1512 f. Facial masking (reduced expression)
 1513 g. Shuffling gait with irregular stride
 1514 2. Non-Motor Early Indicators <citation>id_3, id_5, id_7, id_18</citation>

1512

1513 a. Hyposmia (loss of smell)

1514 b. Constipation

1515 c. Sleep disturbances and REM sleep behavior disorder

1516 d. Depression and anxiety

1517 e. Subtle cognitive changes

1518 f. Autonomic dysfunction (blood pressure fluctuations)

1519 g. Unexplained weight loss

1520 C. Progressive Stage Warning Signs (Stages 2-3)

1521 1. Advancing Motor Symptoms <citation>id_5, id_6, id_18, id_22</citation>

1522 a. Bilateral tremor and rigidity

1523 b. Postural deterioration and stooped posture

1524 c. Balance problems and impaired reflexes

1525 d. Speech changes (soft, monotone voice)

1526 e. Swallowing difficulties

1527 f. Freezing episodes

1528 2. Emerging Complications <citation>id_16, id_18, id_22</citation>

1529 a. Medication wearing-off periods

1530 b. Levodopa-induced dyskinesia

1531 c. Increased fall risk

1532 d. Emotional and mood changes

1533 D. Advanced Stage Warning Signs (Stages 4-5)

1534 1. Severe Motor Impairment <citation>id_5, id_6, id_9, id_22</citation>

1535 a. Major mobility loss requiring assistive devices

1536 b. Severe postural deformities

1537 c. Inability to live independently

1538 d. Wheelchair or bed confinement

1539 2. Complex Non-Motor Complications <citation>id_5, id_16, id_19, id_22</citation>

1540 a. Cognitive decline and dementia

1541 b. Hallucinations and delusions

1542 c. Severe autonomic dysfunction

1543 d. Chronic pain and sexual dysfunction

1544 II. Critical Warning Signs Requiring Immediate Medical Intervention

1545 A. Emergency Situations for Family Members

1546 1. Respiratory and Swallowing Emergencies <citation>id_25, id_26, id_29, id_31, id_32, id_35, id_36</citation>

1547 a. Sudden choking or difficulty swallowing

1548 b. Aspiration risk with coughing while eating

1549 c. Severe voice changes or stridor

1550 d. Respiratory difficulty or shortness of breath

1551 2. Neurological Crisis Signs <citation>id_24, id_26, id_28, id_32, id_33, id_35</citation>

1552 a. Sudden confusion, delirium, or marked cognitive decline

1553 b. New or worsening hallucinations, delusions, or paranoid thoughts

1554 c. Abrupt neurological deficits (weakness, speech changes)

1555 d. Loss of consciousness or seizures

1556 3. Medication-Related Emergencies <citation>id_24, id_28, id_33</citation>

1557 a. Abrupt stopping or alteration of medications

1558 b. Exposure to dopamine-blocking drugs

1559 c. Missed or delayed essential medication doses

1560 d. Signs of serotonin syndrome

1561 B. Cardiovascular and Autonomic Emergencies

1562 1. Circulatory Complications <citation>id_26, id_28, id_31, id_32, id_34, id_36</citation>

1563 a. Severe orthostatic hypotension causing fainting

1564 b. Chest pain or cardiac symptoms

1565 c. Signs of blood clots (leg swelling, pain, redness)

1566 d. Marked dehydration

1567 2. Infection and Systemic Issues <citation>id_24, id_26, id_32, id_35</citation>

1568 a. Sudden fever with respiratory or urinary symptoms

1566
1567 b. Signs of urinary tract infection
1568 c. Post-surgical complications
1569 d. Severe constipation or bowel impaction
1570 C. Psychiatric and Behavioral Emergencies
1571 1. Mental Health Crises <citation>id_28, id_32, id_33, id_36</citation>
1572 a. Severe depression or suicidal thoughts
1573 b. Violent behaviors during REM sleep disorder
1574 c. Severe impulse-control behaviors
1575 d. Acute psychosis with agitation or aggression
1576 2. Device-Related Emergencies <citation>id_28, id_34</citation>
1577 a. Sudden malfunction of implanted devices
1578 b. DBS system problems (infection, battery failure)
1579 c. Unexpected sleep attacks during critical activities
III. Deep Brain Stimulation (DBS) Surgery: Comprehensive Care Framework
1580 A. Pre-Operative Considerations and Evaluation
1581 1. Candidate Selection Criteria <citation>id_45, id_56</citation>
1582 a. Minimum four years of Parkinson's disease
1583 b. Continued medication benefit with motor complications
1584 c. Absence of dementia or severe cognitive impairment
1585 d. Realistic expectations about outcomes
1586 2. Pre-Surgical Assessment Process <citation>id_39, id_45, id_56</citation>
1587 a. Multidisciplinary team evaluation
1588 b. Neurological and neurosurgical consultation
1589 c. Cognitive testing and brain imaging
1590 d. Medication review and optimization
B. Surgical Procedure and Immediate Post-Operative Care
1591 1. Surgical Process <citation>id_45, id_49, id_51, id_56</citation>
1592 a. Electrode implantation in target brain regions
1593 b. Pulse generator placement under collarbone
1594 c. Brief hospital stay (1-2 days)
1595 d. Low mortality rate (<0.5%) and modest complication rates (4-7%)
1596 2. Initial Recovery Phase <citation>id_41, id_47, id_49, id_51, id_52, id_65</citation>
1597 a. Expected post-operative signs (bruising, swelling, tenderness)
1598 b. Temporary "honeymoon" or microlesion effect
1599 c. Activity restrictions (4-6 weeks)
1600 d. Wound care and infection prevention
C. Device Programming and Optimization
1601 1. Programming Timeline <citation>id_41, id_45, id_46, id_47, id_51, id_56</citation>
1602 a. Initial activation 2-4 weeks post-surgery
1603 b. Optimization period of 4-6 months
1604 c. Multiple programming visits required
1605 d. Ongoing adjustments every 6 months
1606 2. Expected Outcomes and Adjustments <citation>id_46, id_51, id_53, id_56</citation>
1607 a. Significant motor symptom improvement (= 32%)
1608 b. Medication reduction (up to 48%)
1609 c. Functional independence gains
1610 d. Possible side effects requiring management
IV. Daily Life Adjustments and Support Strategies for DBS Patients
1611 A. Physical Care and Safety Measures
1612 1. Activity Modifications <citation>id_47, id_51, id_52, id_65, id_67</citation>
1613 a. Gradual resumption of normal activities
1614 b. Continued use of mobility aids to prevent falls
1615 c. Avoidance of high-risk activities (deep water diving, extreme heat)
1616 d. Regular exercise program (minimum 2.5 hours weekly)
1617 2. Device Safety and Maintenance <citation>id_51, id_52, id_69</citation>
1618 a. Protection from electromagnetic interference
1619 b. Carrying handheld controller at all times

1620
 1621 c. Informing healthcare providers about implanted device
 1622 d. Regular battery monitoring and replacement
 1623 B. Psychosocial Support and Adaptation
 1624 1. Identity and Role Adjustment <citation>id_54, id_66, id_68</citation>
 1625 a. Managing "biographical disruption" from rapid symptom relief
 1626 b. Negotiating new caregiver-patient role boundaries
 1627 c. Addressing "burden of normality" phenomenon
 1628 d. Reclaiming personal interests and autonomy
 1629 2. Family Relationship Dynamics <citation>id_54, id_58, id_66, id_68</citation>
 1630 a. Shifting from caregiver-patient to balanced spousal relationships
 1631 b. Reducing over-protective behaviors
 1632 c. Open communication about changing abilities
 1633 d. Professional counseling and cognitive-behavioral therapy
 1634 C. Comprehensive Support Systems
 1635 1. Healthcare Team Coordination <citation>id_39, id_46, id_61, id_72</citation>
 1636 a. Multidisciplinary follow-up care
 1637 b. Specialized Parkinson's nurses
 1638 c. Home-health nursing services
 1639 d. Telemedicine and remote programming options
 1640 2. Family and Caregiver Support <citation>id_58, id_61, id_63, id_66, id_68</citation>
 1641 a. Pre-operative education and realistic expectation setting
 1642 b. Structured psychosocial interventions
 1643 c. Self-help groups and peer support networks
 1644 d. Caregiver self-care and stress management
 1645 D. Practical Daily Living Strategies
 1646 1. Routine and Symptom Management <citation>id_63, id_65, id_67</citation>
 1647 a. Structured daily routines and symptom diaries
 1648 b. Medication schedule adherence
 1649 c. Regular physical activity (walking, tai chi, dancing)
 1650 d. Comfort measures (leg massages, voice training)
 1651 2. Nutritional and Lifestyle Support <citation>id_67</citation>
 1652 a. Mediterranean diet implementation
 1653 b. Adequate rest and sleep hygiene
 1654 c. Intellectual activities and cognitive stimulation
 1655 d. Consultation with specialized healthcare providers
 1656 E. Long-term Monitoring and Adjustment
 1657 1. Ongoing Assessment Needs <citation>id_47, id_52, id_69</citation>
 1658 a. Regular follow-up visits (monthly for first six months)
 1659 b. Symptom tracking and device optimization
 1660 c. Monitoring for complications or side effects
 1661 d. Battery replacement planning (every 3-5 years)
 1662 2. Quality of Life Optimization <citation>id_51, id_54, id_61</citation>
 1663 a. Patient satisfaction monitoring (95% would recommend DBS)
 b. Positive psychology interventions
 c. Character-strength exercises
 d. Long-term relationship and social support maintenance
 </write_outline>

Step 6

<think>

1666 I have now created a comprehensive outline covering all aspects of the user's query. The
 1667 outline includes:
 1668 1. Detailed information about Parkinson's disease stages and warning signs at each stage
 1669 2. Specific emergency warning signs that should alert family members to seek immediate
 1670 medical intervention
 1671 3. Comprehensive information about DBS surgery, including pre-operative considerations,
 1672 surgical process, and post-operative care
 1673 4. Detailed daily life adjustments and support strategies for DBS patients

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

The outline is well-structured with proper hierarchical organization and includes citations from the research materials. However, I should search for any additional information that might be missing, particularly around specific comfort measures and practical tips for families caring for DBS patients.

</think>

<tool_call>

```
{"name": "search", "arguments": {"query": ["Parkinson DBS patient comfort measures home care tips", "DBS device troubleshooting family guide", "Parkinson DBS medication management post surgery"]}, "goal": "Find additional specific comfort measures, troubleshooting guidance, and medication management strategies for families caring for DBS patients"}}
```

</tool_call>

<tool_response>

<id_65>

Summary:

The guide provides comprehensive daily-life support for Parkinson's patients after DBS: keep incisions clean, dry, and allow them to heal without ointments; shower the day after surgery but avoid baths/swimming until the wound check. Resume Parkinson's meds immediately and discuss timing for blood-thinner restart. Limit strenuous activity for 2-4 weeks and incorporate gentle neck stretches to prevent bow-stringing. Expect DBS activation about 4 weeks post-stage 1 and be patient with programming adjustments. Anticipate possible temporary confusion—maintain a consistent routine, strict medication schedule, regular walking, and good sleep (melatonin if needed); arrange caregiver assistance especially if living alone. Monitor for fever, worsening incision signs, facial swelling, new neurological deficits, or worsening confusion, and contact the surgeon promptly. These strategies together offer practical comfort measures and clear family guidance for post-DBS care.

</id_65>

...

</tool_response>

Step 7

<write_outline>

Parkinson's Disease: Health Warning Signs, Family Alert Systems, and Post-DBS Care Strategies

I. Understanding Parkinson's Disease Progression and Warning Signs

A. Disease Staging Systems

1. Hoehn and Yahr Scale Classifications <citation>id_2, id_6, id_9, id_12, id_13, id_14, id_15, id_17, id_20, id_21</citation>

a. Stage 0: No detectable disease

b. Stage 1.0: Unilateral motor symptoms with minimal functional impact

c. Stage 1.5: Unilateral involvement with axial symptoms

d. Stage 2.0: Bilateral symptoms with preserved balance

e. Stage 2.5: Mild bilateral disease with normal pull-test recovery

f. Stage 3.0: Bilateral disease with postural instability but retained independence

g. Stage 4.0: Severe disability requiring assistance but can still walk/stand

h. Stage 5.0: Wheelchair-bound or bedridden without assistance

2. Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS)<citation>id_3</citation>

a. Four-domain assessment approach

b. Motor and non-motor impact evaluation

c. Replacement for older staging systems

3. Disease Progression Timeline and Transition Indicators <citation>id_16, id_17, id_19</citation>

a. Median transition times between stages (20-26 months)

b. HY > 3 as critical threshold for advanced disease

c. Shift from dopaminergic to cortical pathology involvement

B. Early Stage Warning Signs (Stages 1-2)

1. Motor Symptoms <citation>id_2, id_5, id_7, id_18, id_22</citation>

a. Subtle unilateral tremor (often starting in finger)

1728 b. Micrographia (small handwriting)
 1729 c. Bradykinesia (slowed movements)
 1730 d. Mild rigidity and stiffness
 1731 e. Reduced arm swing
 1732 f. Facial masking (reduced expression)
 1733 g. Shuffling gait with irregular stride

1734 2. Non-Motor Early Indicators <citation>id_3, id_5, id_7, id_18</citation>
 1735 a. Hyposmia (loss of smell)
 1736 b. Constipation
 1737 c. Sleep disturbances and REM sleep behavior disorder
 1738 d. Depression and anxiety
 1739 e. Subtle cognitive changes
 1740 f. Autonomic dysfunction (blood pressure fluctuations)
 1741 g. Unexplained weight loss

1742 3. Diagnostic Challenges and Recognition <citation>id_4</citation>
 1743 a. Non-motor signs can appear decades before motor symptoms
 1744 b. Overlap with other disorders complicates diagnosis
 1745 c. Need for asymmetric symptoms or medication response for confirmation

C. Progressive Stage Warning Signs (Stages 2-3)

1746 1. Advancing Motor Symptoms <citation>id_5, id_6, id_18, id_22</citation>
 1747 a. Bilateral tremor and rigidity
 1748 b. Postural deterioration and stooped posture
 1749 c. Balance problems and impaired reflexes
 1750 d. Speech changes (soft, monotone voice)
 1751 e. Swallowing difficulties
 1752 f. Freezing episodes

1753 2. Emerging Complications <citation>id_16, id_18, id_22</citation>
 1754 a. Medication wearing-off periods
 1755 b. Levodopa-induced dyskinesia
 1756 c. Increased fall risk
 1757 d. Emotional and mood changes

1758 3. Functional Impact Assessment <citation>id_6, id_20, id_22</citation>
 1759 a. Loss of independence in daily activities
 1760 b. Need for assistance with dressing, eating, bathing
 1761 c. Occupational and social limitations

D. Advanced Stage Warning Signs (Stages 4-5)

1762 1. Severe Motor Impairment <citation>id_5, id_6, id_9, id_22</citation>
 1763 a. Major mobility loss requiring assistive devices
 1764 b. Severe postural deformities
 1765 c. Inability to live independently
 1766 d. Wheelchair or bed confinement

1767 2. Complex Non-Motor Complications <citation>id_5, id_16, id_19, id_22</citation>
 1768 a. Cognitive decline and dementia
 1769 b. Hallucinations and delusions
 1770 c. Severe autonomic dysfunction
 1771 d. Chronic pain and sexual dysfunction

1772 3. End-Stage Considerations <citation>id_6, id_9, id_20</citation>
 1773 a. Complete dependency for daily activities
 1774 b. Risk of aspiration and respiratory complications
 1775 c. Need for constant supervision and care

II. Critical Warning Signs Requiring Immediate Medical Intervention

A. Emergency Situations for Family Members

1776 1. Respiratory and Swallowing Emergencies <citation>id_25, id_26, id_29, id_31, id_32, id_35, id_36</citation>
 1777 a. Sudden choking or difficulty swallowing
 1778 b. Aspiration risk with coughing while eating
 1779 c. Severe voice changes or stridor

1782 d. Respiratory difficulty or shortness of breath

1783 2. Neurological Crisis Signs <citation>id_24, id_26, id_28, id_32, id_33, id_35</citation>

1784 a. Sudden confusion, delirium, or marked cognitive decline

1785 b. New or worsening hallucinations, delusions, or paranoid thoughts

1786 c. Abrupt neurological deficits (weakness, speech changes)

1787 d. Loss of consciousness or seizures

1788 3. Medication-Related Emergencies <citation>id_24, id_28, id_33</citation>

1789 a. Abrupt stopping or alteration of medications

1790 b. Exposure to dopamine-blocking drugs

1791 c. Missed or delayed essential medication doses

1792 d. Signs of serotonin syndrome

1793 4. Motor System Crisis <citation>id_25, id_28, id_31, id_34</citation>

1794 a. Sudden severe rigidity or immobility

1795 b. Frequent falls or freezing episodes

1796 c. Rapid loss of walking ability

1797 d. Severe dyskinesia with complications

1798 B. Cardiovascular and Autonomic Emergencies

1799 1. Circulatory Complications <citation>id_26, id_28, id_31, id_32, id_34, id_36</citation>

1800 a. Severe orthostatic hypotension causing fainting

1801 b. Chest pain or cardiac symptoms

1802 c. Signs of blood clots (leg swelling, pain, redness)

1803 d. Marked dehydration

1804 2. Infection and Systemic Issues <citation>id_24, id_26, id_32, id_35</citation>

1805 a. Sudden fever with respiratory or urinary symptoms

1806 b. Signs of urinary tract infection

1807 c. Post-surgical complications

1808 d. Severe constipation or bowel impaction

1809 C. Psychiatric and Behavioral Emergencies

1810 1. Mental Health Crises <citation>id_28, id_32, id_33, id_36</citation>

1811 a. Severe depression or suicidal thoughts

1812 b. Violent behaviors during REM sleep disorder

1813 c. Severe impulse-control behaviors

1814 d. Acute psychosis with agitation or aggression

1815 2. Device-Related Emergencies <citation>id_28, id_34</citation>

1816 a. Sudden malfunction of implanted devices

1817 b. DBS system problems (infection, battery failure)

1818 c. Unexpected sleep attacks during critical activities

1819 D. Hospitalization and Perioperative Risks <citation>id_32, id_35</citation>

1820 1. Hospital-Related Complications

1821 a. Medication timing disruptions

1822 b. Delirium and confusion in unfamiliar environments

1823 c. Increased fall risk and injury potential

1824 d. Aspiration pneumonia during procedures

1825 2. Post-Surgical Warning Signs

1826 a. Rapid symptom deterioration

1827 b. Signs of infection or wound complications

1828 c. Unexpected neuropsychiatric changes

1829 III. Deep Brain Stimulation (DBS) Surgery: Comprehensive Care Framework

1830 A. Pre-Operative Considerations and Evaluation

1831 1. Candidate Selection Criteria <citation>id_45, id_56</citation>

1832 a. Minimum four years of Parkinson's disease

1833 b. Continued medication benefit with motor complications

1834 c. Absence of dementia or severe cognitive impairment

1835 d. Realistic expectations about outcomes

1836 2. Pre-Surgical Assessment Process <citation>id_39, id_45, id_56</citation>

1837 a. Multidisciplinary team evaluation

1838 b. Neurological and neurosurgical consultation

1836

1837 c. Cognitive testing and brain imaging

1838 d. Medication review and optimization

1839 3. Target Selection Considerations <citation>id_53</citation>

1840 a. Subthalamic nucleus (STN) for greater medication reduction

1841 b. Globus pallidus interna (GPi) for language/cognition preservation

1842 c. Individual patient factors influencing choice

1843 B. Surgical Procedure and Immediate Post-Operative Care

1844 1. Surgical Process <citation>id_45, id_49, id_51, id_56</citation>

1845 a. Electrode implantation in target brain regions

1846 b. Pulse generator placement under collarbone

1847 c. Brief hospital stay (1-2 days)

1848 d. Low mortality rate (<0.5%) and modest complication rates (4-7%)

1849 2. Initial Recovery Phase <citation>id_41, id_47, id_49, id_51, id_52, id_65</citation>

1850 a. Expected post-operative signs (bruising, swelling, tenderness)

1851 b. Temporary "honeymoon" or microlesion effect

1852 c. Activity restrictions (4-6 weeks)

1853 d. Wound care and infection prevention

1854 3. Early Post-Operative Monitoring <citation>id_65, id_88</citation>

1855 a. Incision care and healing assessment

1856 b. Medication resumption protocols

1857 c. Activity limitation guidelines

1858 d. Signs requiring immediate medical attention

1859 C. Device Programming and Optimization

1860 1. Programming Timeline <citation>id_41, id_45, id_46, id_47, id_51, id_56</citation>

1861 a. Initial activation 2-4 weeks post-surgery

1862 b. Optimization period of 4-6 months

1863 c. Multiple programming visits required

1864 d. Ongoing adjustments every 6 months

1865 2. Expected Outcomes and Adjustments <citation>id_46, id_51, id_53, id_56</citation>

1866 a. Significant motor symptom improvement (= 32%)

1867 b. Medication reduction (up to 48%)

1868 c. Functional independence gains

1869 d. Possible side effects requiring management

1870 3. Programming Process and Patient Education <citation>id_80, id_86, id_90</citation>

1871 a. Patient-controlled adjustments with handheld programmer

1872 b. Battery monitoring and replacement scheduling

1873 c. Understanding normal sensations vs. concerning symptoms

1874 d. Communication protocols with programming team

1875 IV. Daily Life Adjustments and Support Strategies for DBS Patients

1876 A. Physical Care and Safety Measures

1877 1. Activity Modifications <citation>id_47, id_51, id_52, id_65, id_67</citation>

1878 a. Gradual resumption of normal activities

1879 b. Continued use of mobility aids to prevent falls

1880 c. Avoidance of high-risk activities (deep water diving, extreme heat)

1881 d. Regular exercise program (minimum 2.5 hours weekly)

1882 2. Device Safety and Maintenance <citation>id_51, id_52, id_69, id_77, id_88</citation>

1883 a. Protection from electromagnetic interference

1884 b. Carrying handheld controller at all times

1885 c. Informing healthcare providers about implanted device

1886 d. Regular battery monitoring and replacement

1887 3. Environmental Safety Considerations <citation>id_88, id_90</citation>

1888 a. Contraindicated therapies and equipment

1889 b. Safe vs. unsafe daily activities

1889 c. Travel and security considerations

1889 d. Workplace and recreational restrictions

1889 B. Medication Management Post-DBS

1889 1. Immediate Post-Operative Medication Protocol <citation>id_84, id_88, id_93</citation>

1890

- 1891 a. Continuation of pre-surgical regimen initially
- 1892 b. Gradual, systematic reduction approach
- 1893 c. Monitoring for withdrawal symptoms
- 1894 d. Coordination with stimulation programming
- 1895 2. Long-Term Medication Optimization <citation>id_83, id_84, id_85, id_93</citation>
- 1896 a. Levodopa dosing adjustments
- 1897 b. Dopamine agonist tapering strategies
- 1898 c. Management of persistent symptoms
- 1899 d. Monitoring for mood and cognitive changes
- 1900 3. Troubleshooting Medication Issues <citation>id_81, id_86</citation>
- 1901 a. Home-health nurse medication reviews
- 1902 b. Emergency protocols for medication disruption
- 1903 c. Balancing stimulation with pharmaceutical needs
- 1904 d. Communication with neurology team
- 1905 C. Home-Based Care and Comfort Measures
- 1906 1. Daily Routine Optimization <citation>id_63, id_65, id_67, id_72</citation>
- 1907 a. Structured daily routines and symptom diaries
- 1908 b. Sleep hygiene and rest protocols
- 1909 c. Regular physical activity programs
- 1910 d. Comfort interventions (massage, voice training)
- 1911 2. Home Health Services <citation>id_72, id_81</citation>
- 1912 a. Trained home-health nurse visits
- 1913 b. Remote programming capabilities
- 1914 c. Vital sign monitoring and medication adjustments
- 1915 d. Reduced travel burden for families
- 1916 3. Device Troubleshooting at Home <citation>id_77, id_78, id_86, id_90</citation>
- 1917 a. Basic device operation and monitoring
- 1918 b. Recognition of malfunction signs
- 1919 c. Emergency contact protocols
- 1920 d. Battery status management
- 1921 D. Psychosocial Support and Adaptation
- 1922 1. Identity and Role Adjustment <citation>id_54, id_66, id_68</citation>
- 1923 a. Managing "biographical disruption" from rapid symptom relief
- 1924 b. Negotiating new caregiver-patient role boundaries
- 1925 c. Addressing "burden of normality" phenomenon
- 1926 d. Reclaiming personal interests and autonomy
- 1927 2. Family Relationship Dynamics <citation>id_54, id_58, id_66, id_68</citation>
- 1928 a. Shifting from caregiver-patient to balanced spousal relationships
- 1929 b. Reducing over-protective behaviors
- 1930 c. Open communication about changing abilities
- 1931 d. Professional counseling and cognitive-behavioral therapy
- 1932 3. Caregiver Support and Education <citation>id_61, id_66, id_68</citation>
- 1933 a. Pre-operative education and expectation setting
- 1934 b. Structured psychosocial interventions
- 1935 c. Self-help groups and peer support networks
- 1936 d. Caregiver self-care and stress management
- 1937 E. Comprehensive Support Systems
- 1938 1. Healthcare Team Coordination <citation>id_39, id_46, id_61, id_72</citation>
- 1939 a. Multidisciplinary follow-up care
- 1940 b. Specialized Parkinson's nurses
- 1941 c. Home-health nursing services
- 1942 d. Telemedicine and remote programming options
- 1943 2. Long-Term Quality of Life Optimization <citation>id_51, id_54, id_61</citation>
- 1944 a. Patient satisfaction monitoring (95% would recommend DBS)
- 1945 b. Positive psychology interventions
- 1946 c. Character-strength exercises
- 1947 d. Long-term relationship and social support maintenance

1944
 1945 3. Emergency Preparedness and End-of-Life Considerations <citation>id_78</citation>
 1946 a. Off-trial protocols for device assessment
 1947 b. Palliative care coordination
 1948 c. Hardware monitoring in advanced stages
 1949 d. Family decision-making support
 1950 F. Practical Implementation Strategies
 1951 1. Daily Living Skills <citation>id_60, id_63, id_67</citation>
 1952 a. Gradual re-engagement in routine tasks
 1953 b. Nutritional support (Mediterranean diet)
 1954 c. Intellectual and social activities
 1955 d. Family encouragement and participation
 1956 2. Long-term Monitoring and Adjustment <citation>id_47, id_52, id_69</citation>
 1957 a. Regular follow-up visits (monthly for first six months)
 1958 b. Symptom tracking and device optimization
 1959 c. Monitoring for complications or side effects
 1960 d. Battery replacement planning (every 3-5 years)
 1961 3. Resource Utilization and Support Networks <citation>id_55, id_58, id_76, id_79</citation>
 1962 a. Educational materials and patient guides
 1963 b. Webinars and professional consultations
 1964 c. Support groups and peer networks
 1965 d. Specialized healthcare provider access
 1966 </write_outline>
 1967 **Step 8**
 1968 <think>
 1969 I have now created a comprehensive, detailed outline that addresses all aspects of the user's
 1970 query. The outline includes:
 1971 1. Detailed Parkinson's disease staging and progression with comprehensive warning signs
 1972 2. Specific emergency situations requiring immediate medical intervention
 1973 3. Complete DBS surgery information including pre-operative, operative, and post-operative
 1974 care
 1975 4. Extensive daily life adjustments and support strategies for DBS patients
 1976 5. Proper citations for all subsections
 1977 The outline is well-structured with four levels of hierarchy and covers all essential
 1978 aspects requested. It provides practical, actionable information for families caring for
 1979 Parkinson's patients, particularly those who have undergone DBS surgery. The outline is
 1980 comprehensive, logically organized, and includes insights into the reasoning behind various
 1981 care strategies.
 1982 I believe this outline now fully addresses the user's comprehensive query about Parkinson's
 1983 disease warning signs, family alert systems, and post-DBS care strategies.
 1984 <think>
 1985 <terminate> ✓

I.7 CASE STUDY OF WRITING

We present only the first 7 steps of the long writing trajectory to save space. The afterwards reasoning process is similar to that in the first 7 steps.

1992 Case Trajectory of Writing in DeepResearch Bench
 1993
 1994 **Question:** Parkinson's disease has a profound impact on patients. What are the potential
 1995 health warning signs associated with different stages of the disease? As family members,
 1996 which specific signs should alert us to intervene or seek medical advice regarding the
 1997 patient's condition? Furthermore, for patients who have undergone Deep Brain Stimulation

1998
 1999 (DBS) surgery, what daily life adjustments and support strategies can be implemented to
 2000 improve their comfort and overall well-being?

2001 **Step 1**

2002 **<think>**
 2003 I need to write a comprehensive article about Parkinson's disease warning signs, emergency
 2004 situations for family members, and post-DBS care strategies. Looking at the outline, I should
 2005 start by retrieving information about disease staging systems and warning signs. Let me begin
 2006 with the first section on understanding Parkinson's disease progression and warning signs.

2007 **</think>**
 2008 **<tool_call>**
 2009 {"name": "retrieve", "arguments": {"url_id": ["id_2", "id_6", "id_9", "id_12", "id_13",
 2010 "id_14", "id_15", "id_17", "id_20", "id_21"], "goal": "Gather comprehensive information
 2011 about Hoehn and Yahr scale classifications and disease staging systems for Parkinson's
 2012 disease"}}
 2013 **</tool_call>**
 2014 **<tool_response>**
 2015 The page content for the previous section has been masked for saving the space.
 2016 **</tool_response>**

2017 **Step 2**

2018 **<think>**
 2019 I have comprehensive information about the Hoehn and Yahr scale classifications and disease
 2020 staging systems. This provides a solid foundation for understanding Parkinson's disease
 2021 progression. I can see the clear progression from Stage 0 (no disease) through Stage 5
 2022 (wheelchair/bedridden), with intermediate stages 1.5 and 2.5 in the modified version. The
 2023 information shows both motor and non-motor symptoms, timing of transitions between
 2024 stages, and functional impacts. I should now write the takeaway section and begin the first
 2025 major section on disease staging systems.

2026 **</think>**
 2027 **<write>**
 2028 **<takeaway>**
 2029 Parkinson's disease progresses through distinct stages with recognizable warning signs that
 2030 families must understand to provide appropriate care and seek timely medical intervention.
 2031 The disease advances from subtle unilateral symptoms in early stages to severe bilateral
 2032 impairment requiring constant care in advanced stages. Critical emergency signs include
 2033 sudden choking, severe falls, medication disruptions, neurological changes, and psychiatric
 2034 crises that demand immediate medical attention. For patients who have undergone Deep
 2035 Brain Stimulation (DBS) surgery, comprehensive daily life adjustments encompass structured
 2036 post-operative care, medication management, device safety protocols, psychosocial
 2037 support, and coordinated family involvement to optimize comfort and long-term outcomes.
 2038 Understanding these progressive warning signs and implementing evidence-based support
 2039 strategies enables families to navigate the complex challenges of Parkinson's disease while
 2040 maintaining patient dignity and quality of life.
 2041 **</takeaway>**
 2042 Parkinson's Disease: Health Warning Signs, Family Alert Systems, and Post-DBS Care
 2043 Strategies
 2044 I. Understanding Parkinson's Disease Progression and Warning Signs
 2045 A. Disease Staging Systems
 2046 Understanding Parkinson's disease progression requires familiarity with established staging
 2047 systems that provide frameworks for recognizing symptom evolution and functional decline.
 2048 These systems serve as critical tools for families to understand what to expect and when to
 2049 seek medical intervention.
 2050 1. Hoehn and Yahr Scale Classifications
 2051 The Hoehn and Yahr scale, originally published in 1967, remains the most widely used
 2052 system for staging Parkinson's disease progression. **<cite id="id_14,id_17">**This scale
 2053 distinguishes between five stages of Parkinson's Disease, from unilateral impairment (Stage
 2054 1) to bilateral impairment, without postural control difficulties (Stage 2), to postural instability

2052
 2053 (Stage 3), to loss of physical independence (Stage 4) and finally, to wheelchair-bound or
 2054 bed-bound (Stage 5)</cite>. The modified version includes intermediate stages that capture
 2055 more nuanced progression patterns.
 2056 <table>
 2057 <caption>Hoehn and Yahr Scale Progression and Warning Signs</caption>
 2058 <thead>
 2059 <tr>
 2060 <th>Stage</th>
 2061 <th>Primary Characteristics</th>
 2062 <th>Functional Impact</th>
 2063 <th>Key Warning Signs for Families</th>
 2064 </tr>
 2065 </thead>
 2066 <tbody>
 2067 <tr>
 2068 <td>Stage 0</td>
 2069 <td><cite id="id_12,id_21">No signs of disease</cite></td>
 2070 <td>Normal function</td>
 2071 <td>Baseline monitoring period</td>
 2072 </tr>
 2073 <tr>
 2074 <td>Stage 1.0</td>
 2075 <td><cite id="id_15,id_20">Unilateral tremor, rigidity, reduced arm swing, and slowness
 2076 present only on one side</cite></td>
 2077 <td><cite id="id_2,id_6">Minimal interference with daily activities</cite></td>
 2078 <td>Subtle tremor in one finger, asymmetric arm swing</td>
 2079 </tr>
 2080 <tr>
 2081 <td>Stage 1.5</td>
 2082 <td><cite id="id_15,id_21">Unilateral involvement with axial symptoms like reduced facial
 2083 expression, stooped posture, or reduced arm swing on both sides</cite></td>
 2084 <td>Slight postural changes</td>
 2085 <td>Facial masking, posture changes visible to family</td>
 2086 </tr>
 2087 <tr>
 2088 <td>Stage 2.0</td>
 2089 <td><cite id="id_13,id_17">Bilateral involvement without impairment of bal-
 2090 ance</cite></td>
 2091 <td><cite id="id_2,id_6">Daily activities become more difficult but independent living
 2092 realistic</cite></td>
 2093 <td>Tremor and stiffness on both sides, speech difficulties</td>
 2094 </tr>
 2095 <tr>
 2096 <td>Stage 2.5</td>
 2097 <td><cite id="id_15,id_21">Mild bilateral disease with recovery on pull test (taking ≥ 3
 2098 steps to recover balance)</cite></td>
 2099 <td>Mild balance impairment</td>
 2100 <td>Slight unsteadiness, requires multiple steps to recover balance</td>
 2101 </tr> <tr> <td>Stage 3.0</td>
 2102 <td><cite id="id_13,id_20">Mild to moderate bilateral disease with impaired postural
 2103 reflexes; physically independent</cite></td>
 2104 <td><cite id="id_2,id_6">Falls become frequent, assistance needed for some activi-
 2105 ties</cite></td> <td>Loss of balance when pulled backward, increased fall risk</td>
 </tr> <tr> <td>Stage 4.0</td>
 <td><cite id="id_13,id_20">Severe disability; still able to walk or stand unassisted but
 markedly incapacitated</cite></td>
 <td><cite id="id_2,id_6">Cannot live independently, requires walker or assis-
 </td>

```

2106
2107 tance</cite></td> <td>Need for assistive devices, inability to live alone safely</td>
2108 </tr> <tr> <td>Stage 5.0</td>
2109 <td><cite id="id_13,id_21">Wheelchair bound or bedridden unless aided</cite></td>
2110 <td><cite id="id_2,id_6">Around-the-clock nursing care required, up to 50% experience
2111 hallucinations and delusions</cite></td>
2112 <td>Inability to stand or walk, cognitive disturbances, complete dependency</td> </tr>
2113 </tbody>
2114 </table>
2115 This staging system reveals several critical insights for families. **The transition from
2116 Stage 2 to Stage 3 marks a pivotal point** where balance problems emerge, fundamentally
2117 changing the patient's safety profile and care needs. <cite id="id_20">Loss of balance, with
2118 the inability to make the rapid, automatic and involuntary movements necessary to protect
2119 against falling, is one of the most troubling and dangerous aspects of Parkinsonism and one
2120 of the least easily treated</cite>. Understanding this progression helps families anticipate
2121 when environmental modifications and increased supervision become essential.
2122 **The distinction between Stage 3 and Stage 4 centers on independence**. <cite
2123 id="id_20">The most important factor identifying Stage III (as opposed to stage IV) is that
2124 the patient is still fully independent in all activities of daily living (dressing, hygiene, eating,
2125 etc.), while Stage IV patients are unable to lead an independent life because of the need for
2126 help with some activities of daily living</cite>. This transition signals when families must
2127 consider more intensive caregiving arrangements or assisted living options.
2128 </write>
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159

```

Step 3

<think>

Now I need to continue with information about the MDS-UPDRS system and disease progression timeline. Let me retrieve more specific information about these aspects.

</think>

<tool_call>

```
{"name": "search", "arguments": { "url_id": ["id_3", "id_16", "id_17", "id_19"], "goal": "Gather information about MDS-UPDRS assessment system, disease progression timeline, and transition indicators between stages"}}
```

</tool_call>

<tool_response>

The page content for the previous section has been masked for saving the space.

</tool_response>

Step 4

<think>

The tool response provides valuable information about the MDS-UPDRS system, disease progression timeline, and the critical transition point at Stage 3. I can see that:

1. MDS-UPDRS has replaced the older Hoehn-Yahr system as the main assessment tool
2. The progression timeline shows median transition times between stages (20-26 months typically)
3. Stage 3 (HY > 3) represents a critical threshold marking advanced disease with increased disability, medication-refractory features, and cognitive decline
4. There's important information about non-motor symptoms appearing years before motor symptoms
5. Specific progression warning signs are detailed

I should continue writing about the MDS-UPDRS system and progression timeline, then move on to early stage warning signs.

</think>

<tool_call>

```
{"name": "search", "arguments": ["DBS patient daily living support family caregivers", "Parkinson DBS lifestyle modifications comfort strategies", "DBS programming adjustment patient family guide"], "goal": "Find specific daily life support strategies, comfort measures, and family guidance for caring for Parkinson's patients who have undergone DBS surgery"}}
```

</tool_call>

<write>

2160

2. Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS)
 While the Hoehn and Yahr scale remains valuable for understanding disease progression, *<cite id="id_3">*the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is healthcare providers' main tool to classify this disease. The MDS-UPDRS examines four different areas of how Parkinson's disease affects you*</cite>*. This comprehensive assessment tool provides a more nuanced evaluation of both motor and non-motor symptoms, addressing the limitations of purely stage-based approaches.

The significance of this shift toward individualized assessment cannot be overstated. Rather than simply categorizing patients into stages, the MDS-UPDRS recognizes that Parkinson's disease affects each person differently, requiring personalized treatment approaches. This evolution in assessment methodology reflects our growing understanding that the disease's impact extends far beyond motor symptoms.

3. Disease Progression Timeline and Transition Indicators

Understanding the temporal aspects of Parkinson's progression helps families prepare for changing care needs. *<cite id="id_17">*Median transition times between stages show Stage 2 progression occurring around 20 months, with subsequent stages progressing every 24-26 months*</cite>*. However, *<cite id="id_16">*for some people, it can take as long as 20 years to move through all the stages*</cite>*, emphasizing the highly variable nature of disease progression.

The transition to Stage 3 represents a critical threshold in Parkinson's disease progression. *<cite id="id_19">*Transitions from scores of ≤ 2.5 to scores of ≥ 3 on the modified HY scale mark a worsening in quality of life and an increase in the risk of medication-refractory disease features such as dementia and mortality*</cite>*. This transition point signals when families must significantly intensify their care planning and support strategies.

The underlying pathophysiology also changes at this critical juncture. *<cite id="id_19">*Although striatal dopaminergic neuropathology accounts for many of the early motor features of PD, *in vivo* imaging studies suggest floor effects to striatal dopaminergic denervation that occur both in early PD and at HY stage 3 disease, at which point progressive disease burden is driven by other factors*</cite>*. This shift from primarily dopaminergic to multi-system pathology explains why treatment becomes more complex and why families observe new types of symptoms emerging.

B. Early Stage Warning Signs (Stages 1-2)

Early recognition of Parkinson's disease symptoms enables timely medical intervention and better long-term outcomes. The challenge for families lies in distinguishing between normal aging changes and early disease manifestations.

1. Motor Symptoms

The classic motor symptoms of Parkinson's disease typically appear gradually and asymmetrically. *<cite id="id_3">*Motor symptoms — which means movement-related symptoms — of Parkinson's disease include tremor at rest, bradykinesia (slowness of movement), rigidity (muscle stiffness), and postural instability (balance problems)*</cite>*.

Tremor characteristics provide important diagnostic clues. The typical Parkinson's tremor occurs at rest, often beginning subtly in one finger or hand. Families should note that this tremor typically diminishes during purposeful movement, distinguishing it from other types of tremors. The asymmetric onset is particularly significant—symptoms beginning on one side of the body are more suggestive of Parkinson's disease than bilateral symptoms.

Bradykinesia manifests in multiple ways that families can observe in daily activities. This slowness of movement affects not only gross motor activities but also fine motor skills, leading to characteristic changes in handwriting (micrographia), reduced facial expression (facial masking), and decreased arm swing during walking. These changes often develop so gradually that they may be attributed to normal aging until they become more pronounced.

Rigidity and postural changes represent another category of early motor symptoms. Families may notice increased muscle stiffness, particularly in the neck and shoulders, leading to complaints of aching or discomfort. The development of a stooped posture and reduced arm swing during walking are early indicators that warrant medical evaluation.

2. Non-Motor Early Indicators

Perhaps most significantly for early detection, *<cite id="id_3">*many of the possible non-motor symptoms can appear years or even decades ahead of motor symptoms*</cite>*.

2214
 2215 This extended prodromal period offers families crucial opportunities for early intervention,
 2216 though the subtlety of these symptoms often leads to delayed recognition.
 2217 ****Olfactory dysfunction represents one of the earliest detectable changes**.** *<cite
 id="id_3">Loss of sense of smell (anosmia)</cite>* occurs in approximately 90% of
 2218 Parkinson's patients, often years before motor symptoms appear. Families should take note
 2219 when a loved one reports that food no longer tastes the same or when they fail to detect
 2220 strong odors that others notice readily.
 2221 ****Gastrointestinal symptoms, particularly constipation****, frequently precede motor
 2222 symptoms by years. *<cite id="id_3">Constipation</cite>* in Parkinson's disease results from
 2223 the same pathological processes affecting the brain, specifically involving the enteric nervous
 2224 system. This symptom is often dismissed as age-related or dietary, but persistent constipation
 2225 without clear cause warrants medical attention.
 2226 ****Sleep disturbances provide another early warning system**.** *<cite id="id_3">Rapid
 eye movement sleep behavior disorder</cite>* is particularly significant, as it involves
 2227 acting out dreams during sleep, potentially including violent movements or vocalizations.
 2228 *<cite id="id_16">REM sleep behavior disorder can also happen in people who don't have
 Parkinson's. However, if this isn't something you've dealt with before, it's likely related to
 your disease</cite>*. Family members sharing a bed are often the first to notice these changes.
 2229 ****Neuropsychiatric symptoms deserve particular attention**** from families. *<cite
 id="id_3">Depression or anxiety</cite>* can appear years before motor symptoms, and *<cite
 id="id_16">anxiety and depression have been linked to Parkinson's. In addition to movement
 problems, the disease can also have an impact on your mental health</cite>*. These mood
 2230 changes may be subtle initially but can significantly impact quality of life and relationships.
 2231
 2232 **</write>**

Step 5

2233
 2234 **<think>**
 2235 Now I need to continue with information about progressive stage warning signs and advanced
 2236 stage warning signs. Let me retrieve more specific information about these later stages and
 2237 their complications.
 2238

2239 **</think>**
 2240 **<tool_call>**
 2241 *{"name": "retrieve", "arguments": { "url_id": ["id_5", "id_6", "id_18", "id_22"], "goal":
 "Gather detailed information about progressive and advanced stage warning signs, including
 motor and non-motor complications"}}*
 2242 **</tool_call>**

2243 **<tool_response>**
 2244 The page content for the previous section has been masked for saving the space.
 2245 **</tool_response>**

Step 6

2246
 2247 **<think>**
 2248 The tool response provides excellent detailed information about progressive and advanced
 2249 stage warning signs. I can see clear patterns in the progression from mild to moderate to
 2250 advanced stages, with specific motor and non-motor symptoms at each stage. The information
 2251 shows how symptoms evolve from unilateral to bilateral, how functional independence
 2252 is progressively lost, and how non-motor symptoms become increasingly prominent and
 2253 disabling. I should organize this information to show the progression clearly and highlight
 2254 the key warning signs families need to recognize.
 2255

2256 **</think>**
 2257 **<write>**
 2258 C. Progressive Stage Warning Signs (Stages 2-3)
 2259 As Parkinson's disease advances beyond the early stage, families observe more pronounced
 2260 symptoms that begin to significantly impact daily functioning. This progressive phase
 2261 represents a critical period where proactive intervention can substantially influence quality of
 2262 life outcomes.
 2263

1. Advancing Motor Symptoms

2264 ****The bilateral spread of symptoms marks a fundamental shift**** in disease progression.
 2265

2268

2269

2270

2271

2272

2273

2274

2275

2276

2277

2278

2279

2280

2281

2282

2283

2284

2285

2286

2287

2288

2289

2290

2291

2292

2293

2294

2295

2296

2297

2298

2299

2300

2301

2302

2303

2304

2305

2306

2307

2308

2309

2310

2311

2312

2313

2314

2315

2316

2317

2318

2319

2320

2321

<cite id="id_18">As the disease progresses, it eventually affects both sides. However, the symptoms may still be more severe on one side than on the other</cite>. This bilateral involvement signals that families must adapt their support strategies to address more comprehensive functional limitations.

<cite id="id_22">Often within 3 to 7 years, you'll see more changes. Early on, you might have a little trouble with something like buttoning a shirt. At this point, you may not be able to do it at all</cite>. **This progression from mild difficulty to complete inability** represents a pattern families will observe across multiple daily activities, requiring gradual assumption of caregiving responsibilities.

Speech and swallowing changes become particularly noticeable during this stage. <cite id="id_5">You may speak softly or quickly, slur, or hesitate before talking. Your speech may be flat or monotone, without typical speech patterns</cite>. These changes affect not only communication but also safety, as <cite id="id_22">trouble swallowing</cite> increases the risk of aspiration and nutritional deficiencies.

Gait disturbances and freezing episodes represent some of the most challenging aspects of progressive Parkinson's disease. <cite id="id_18">People with Parkinson's disease often develop a parkinsonian gait that includes a tendency to lean forward; take small, quick steps; and reduce swinging their arms. They also may have trouble initiating or continuing movement</cite>. The emergence of <cite id="id_22">freezing when you first start to walk or change direction, as if your feet are glued to the ground</cite> creates significant safety concerns and mobility limitations.

2. Emerging Complications

Medication effectiveness begins to fluctuate during the progressive stage, creating new challenges for symptom management. <cite id="id_22">You might also find that the medicine you take starts to wear off between doses</cite>. This "wearing-off" phenomenon requires families to become more attentive to timing of medications and to recognize when symptoms return between doses.

Balance problems and fall risk increase substantially. <cite id="id_6">At this stage, falls become more frequent due to balance and reflex problems</cite>, and <cite id="id_22">falls to be more likely</cite>. This increased fall risk necessitates environmental modifications and potentially the introduction of assistive devices.

3. Functional Impact Assessment

The distinction between Stages 2 and 3 centers on independence preservation. <cite id="id_6">Due to the various issues this stage presents, daily tasks are much more difficult to do, but people can still live independently. A combination of medication and therapy can help to manage the symptoms mentioned above</cite>. Understanding this threshold helps families recognize when to increase support while still encouraging independence.

D. Advanced Stage Warning Signs (Stages 4-5)

The advanced stages of Parkinson's disease present complex challenges that require comprehensive care planning and family coordination. These stages are characterized by severe functional limitations and the emergence of complications that significantly impact quality of life.

1. Severe Motor Impairment

Stage 4 marks the loss of independent living capability. <cite id="id_6">The critical factor in separating people with stage 3 Parkinson's and stage 4 is independence. Motor skills and deep brain stimulation are heavily impacted at stage 4 and as such, movement symptoms affect a person's ability to retain their independence</cite>. This transition requires families to make difficult decisions about living arrangements and care provision.

<cite id="id_22">At this point, you likely: Are limited to bed or a wheelchair, Can't live on your own, Have severe posture issues in your neck, back, and hips, Need help with daily tasks</cite>. **The comprehensive nature of these limitations** means that families must coordinate multiple aspects of care, from basic hygiene to mobility assistance.

Stage 5 represents the most severe manifestation of motor symptoms. <cite id="id_6">Severe stiffness can make it difficult, if not impossible, for a person to stand or walk. This is due to stiffness causing the legs to essentially freeze when the patient attempts to stand</cite>. This "freezing" phenomenon at the advanced stage differs qualitatively from earlier freezing episodes, representing a more permanent limitation rather than intermittent

2322

2323 difficulty.

2324 2. Complex Non-Motor Complications

2325 ****Cognitive decline becomes a prominent feature**** in advanced stages. [<cite id="id_18">](#)Over time, as the disease progresses, some people may develop dementia and be diagnosed with Parkinson's dementia, a type of Lewy body dementia. People with Parkinson's dementia may have severe memory and thinking problems that affect daily living</cite>. This cognitive impairment adds another layer of complexity to caregiving responsibilities.2330 ****Psychiatric symptoms emerge as significant challenges****. [<cite id="id_6">](#)Furthermore, up to 50% of Parkinson's patients in stages 4 and 5 experience confusion, hallucinations, and delusions. For clarity, hallucinations are seeing things that aren't really there, and delusions are when a person believes something despite evidence to the contrary</cite>. These symptoms can be particularly distressing for families and may require specialized psychiatric intervention.2335 ****Swallowing and nutritional complications**** become life-threatening concerns. [<cite id="id_5">](#)Late-stage Parkinson's disease affects the muscles in the mouth. This causes trouble swallowing and chewing, which can lead to not getting enough nutrients in your diet. If food or saliva collects in the mouth, it can cause choking or drooling</cite>. These complications require careful monitoring and may necessitate alternative feeding methods.

2340 <table>

Stage	Motor Warning Signs	Non-Motor Warning Signs	Functional Impact	Family Action Required
Early (1-2)	Unilateral tremor, mild rigidity, reduced arm swing, micrographia	Sleep disturbances, constipation, loss of smell, mood changes	Minimal daily life disruption	Medical evaluation, symptom monitoring
Progressive (2-3)	Bilateral symptoms, gait freezing, speech changes, balance problems	Medication wearing-off, cognitive changes, swallowing difficulty	Increased difficulty with daily tasks, still independent	Environmental modifications, fall prevention, medication timing
Advanced (4-5)	Severe rigidity, wheelchair/bed bound, postural deformities	Dementia, hallucinations, severe dysphagia, autonomic failure	Complete dependency for daily activities	24-hour care, safety monitoring, end-of-life planning

2363 ****The temporal progression of non-motor symptoms**** follows a predictable pattern that families should understand. [<cite id="id_22">](#)What may show up early. You may have these issues years before any classic motor symptoms like tremors: Constipation, Depression, Loss of smell, Low blood pressure when you stand up, Pain, Sleep issues</cite>. Conversely, [<cite id="id_22">](#)What may show up later. Dementia and psychosis are two serious mental health issues that usually take a while to show up</cite>.2369 This progression pattern reveals that ****non-motor symptoms often become more disabling than motor symptoms**** in advanced stages. [<cite id="id_22">](#)Almost everyone with Parkinson's gets at least one of these. When severe, they're more likely than motor issues to lead to a disability or make you move into a nursing home</cite>. This insight helps families understand why comprehensive care planning must address both motor and non-motor aspects of the disease.

2374 3. End-Stage Considerations

2375

2376
2377 **The advanced stage requires families to confront difficult realities** about prognosis and
2378 care goals. <cite id="id_22">After about 10 years, most people will have at least one major
2379 issue, like dementia or a physical disability</cite>. Understanding this timeline helps families
2380 engage in advance care planning while the patient can still participate in decision-making.
2381 **The variability in progression** means that some individuals may never reach the most
2382 severe stages, while others progress more rapidly. <cite id="id_22">Some people never reach
2383 this stage</cite> of advanced disability, emphasizing the importance of individualized care
2384 planning rather than assuming inevitable decline. </write>

Step 7

<think>

Now I need to move on to the second major section about critical warning signs requiring immediate medical intervention. This is crucial information for families to know when to seek emergency care. Let me retrieve information about emergency situations and warning signs. </think>

<tool call>

```
{"name": "retrieve", "arguments": {"url_id": ["id_24", "id_25", "id_26", "id_28", "id_29", "id_31", "id_32", "id_33", "id_34", "id_35", "id_36"], "goal": "Gather comprehensive information about emergency warning signs and situations that require immediate medical intervention for Parkinson's patients"}}
```

</tool_call>

<tool_response>

The page content for the previous section has been masked for saving the space.

</tool_response>

Step 8

111

<terminate> ✓

2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429