# DR-Arena: an Automated Evaluation Framework for Deep Research Agents

**Yiwen Gao**[1]* **Ruochen Zhao**[2]* **Yang Deng**[3] **Wenxuan Zhang**[4]†
[1]National University of Singapore [2]Nanyang Technological University
[3]Singapore Management University [4]Singapore University of Technology and Design
yiwen_gao@u.nus.edu ruochen002@e.ntu.edu.sg
ydeng@smu.edu.sg wxzhang@sutd.edu.sg

https://github.com/iNLP-Lab/DR-Arena

## Abstract

As Large Language Models (LLMs) increasingly operate as **Deep Research (DR) Agents** capable of autonomous investigation and information synthesis, reliable evaluation of their task performance has become a critical bottleneck. Current benchmarks predominantly rely on static datasets, which suffer from several limitations: *limited task generality*, *temporal misalignment*, and *data contamination*. To address these, we introduce **DR-Arena**, a fully automated evaluation framework that pushes DR agents to their capability limits through dynamic investigation. DR-Arena constructs real-time Information Trees from fresh web trends to ensure the evaluation rubric is synchronized with the live world state, and employs an automated Examiner to generate structured tasks testing two orthogonal capabilities: *Deep reasoning* and *Wide coverage*. DR-Arena further adopts Adaptive Evolvement Loop, a state-machine controller that dynamically escalates task complexity based on real-time performance, demanding deeper deduction or wider aggregation until a decisive capability boundary emerges. Experiments with six advanced DR agents demonstrate that DR-Arena achieves a Spearman correlation of 0.94 with the LMSYS Search Arena leaderboard. This represents state-of-the-art alignment with human preferences without any manual efforts, validating DR-Arena as a reliable alternative for costly human adjudication.

## 1 Introduction

Deep Research (DR) agents, such as OpenAI Deep Research (OpenAI, 2025) and Perplexity Deep Research (Perplexity AI, 2025), have rapidly gained adoption and are now widely used for complex information-seeking tasks. Unlike traditional search engines where users need to browse multiple websites manually, DR agents act as autonomous research agents that conduct multi-step investigations over extended horizons, iteratively retrieving, cross-referencing, and synthesizing evidence from the live web to produce structured and citation-backed reports (Nakano et al., 2022; Shao et al., 2024; Qin et al., 2023). As DR agents are increasingly deployed in real-world and high-stakes analytical settings, efficient and reliable evaluation of their capabilities has become a pressing challenge.

Recent efforts have begun to evaluate deep research capabilities by constructing dedicated datasets for multi-step web-based investigation (Mialon et al., 2024; Wong et al., 2025; Lan et al., 2025). While these benchmarks provide useful performance indicators, they exhibit three fundamental limitations: 1) **limited task generality**, as task-centric and aspect-specific dataset construction restricts evaluation to predefined investigation patterns and weakens transferability to real-world research settings; 2) **temporal misalignment**, since static benchmarks inevitably decay as underlying facts evolve, resulting in evaluation against outdated ground truth; and 3) **data contamination**, whereby persistent and widely reused datasets increasingly appear in model training corpora, resulting in parametric memorization rather than genuine reasoning and evidence synthesis (Ravaut et al., 2025). Collectively, these issues reflect a structural mismatch with DR agents, which are designed to operate in open evolving environments that require adaptive exploration and reasoning.

To address these challenges, we propose DR-Arena, a fully automated evaluation framework that simulates a competitive arena, where DR agents are pushed to their capability limits through dynamic investigation. As shown in Figure 1, DR-Arena operates as a closed-loop framework orchestrated by an automated Examiner, which functions as an interviewer role that drafts questions, assesses candidate answers, and conducts follow-up rounds. To draft dynamic questions, the Examiner first con-

* Equal contribution.
† Corresponding author.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27130–27152

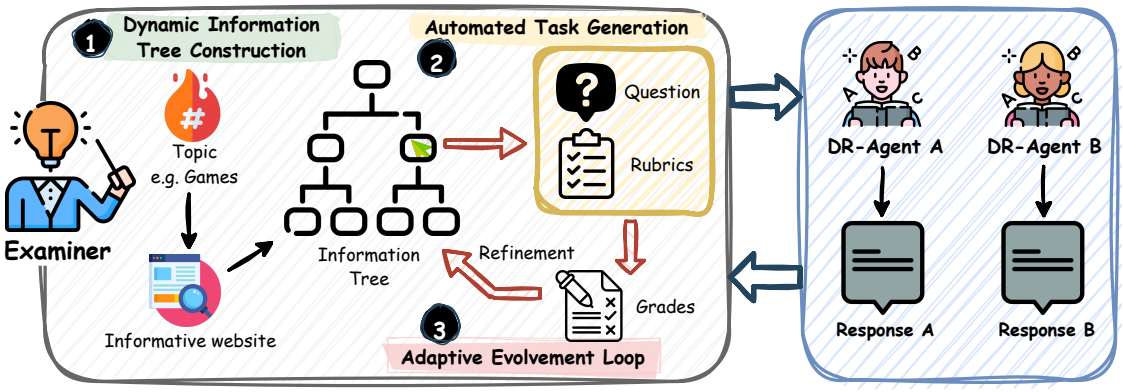


Figure 1: Overview of the DR-Arena Framework.

structs Information Trees by scraping informative websites in real-time. Based on the trees, the Examiner then devises a challenging question for the candidate DR agents. Since the Examiner has access to the ground-truth information source, it also keeps a mental set of grounded rubrics. After the agents respond, the Examiner assesses their performances based on the rubrics. Finally, it enters the Adaptive Evolvement Loop, where the Examiner dynamically decides whether to further refine the tree and ask follow-up questions.

During Adaptive Evolvement, DR-Arena stress-tests two fundamental dimensions: multi-hop reasoning (*Depth*) and information gathering (*Width*). The Examiner actively intervenes based on real-time performance to amplify differences between candidates: When agents reach a stalemate, the Examiner first diagnoses the failure type, whether the mistake originates from a retrieval gap or reasoning deficit, and then intentionally probes that specific weakness in subsequent rounds. This dynamic investigation, guided by the *Depth* and *Width* principles (Lan et al., 2025), ensures that both core abilities are tested comprehensively, pushing DR agents to their capability boundaries.

Experiments across six state-of-the-art DR agents show that DR-Arena achieves superior alignment with human preferences compared to existing benchmarks, accurately recovering the hierarchy of the LMSYS Search Arena (Miroyan et al., 2026) even without manual annotations. Beyond alignment, we demonstrate that the adaptive loop optimizes computational allocation, concentrating computational resources on distinguishing closely matched models while rapidly resolving mismatches. Furthermore, we conduct extensive analysis that reveals distinct cognitive profiles among top-tier DR agents, distinguishing models with robust, balanced capabilities from those exhibiting asymmetric trade-offs between logic and coverage. In conclusion, DR-Arena serves as a trustworthy and scalable proxy for human adjudication, offering the community a reliable standard to benchmark the next generation of DR agents.

## 2 Related Work

**Static Benchmarks for Deep Research.** Evaluation in Agentic Search has traditionally relied on static benchmarks. Early works like *WebArena* (Zhou et al., 2024) and *BrowseComp* (Wei et al., 2025) focused on transactional web tasks, while later benchmarks extended to deep research via expert-curated datasets, including *DeepResearch Bench* (Du et al., 2025) and *DeepResearch Arena* (Wan et al., 2026). Despite increased task complexity, these benchmarks remain static and thus suffer from *temporal degradation* (Kasai et al., 2023): As real-world information evolves, fixed gold answers become outdated, penalizing agents that correctly retrieve fresh evidence (Vu et al., 2024). Approaches such as *LiveBench* (White et al., 2025) mitigate contamination through periodic updates, but still rely on batch refreshes rather than real-time generation. DR-Arena fundamentally departs from this paradigm by generating Dynamic Information Trees directly from the live web, synchronizing tasks and ground truth with the current world state and eliminating both temporal misalignment and parametric memorization.

**Automated Evaluation Frameworks.** To scale beyond human review, prior work has adopted automated evaluation frameworks, most notably *LLM-as-a-Judge* (Zheng et al., 2023) and its refinements in leaderboards such as *AlpacaEval* (Dubois et al., 2025) and *Auto-Arena* (Zhao et al., 2025), which re-

duce single-judge bias via multi-agent deliberation with "Peer Battle" and "Committee" mechanisms. In the agentic domain, frameworks like *Mind2Web-2* (Gou et al., 2025) and *Auto-Eval Judge* (Bhonsle et al., 2025) have evolved into "Agent-as-a-Judge" systems, utilizing structured checklists to verify step-by-step execution. However, the majority of these frameworks operate as passive evaluators, where they grade static trajectories after the task is completed. While *FACT-AUDIT* (Lin et al., 2025) introduces adaptive stress-testing, it remains limited to short-horizon factual claims. DR-Arena departs from this paradigm by introducing an active Examiner for long-horizon Deep Research, where an *Adaptive Evolvement Loop* dynamically modulates task *depth* and *width* during evaluation to expose capability boundaries that static datasets or passive evaluators fail to expose.

## 3 DR-Arena Framework

To probe the limits of DR agents, DR-Arena operates as a closed-loop system orchestrated by a unified LLM agent, denoted as Examiner. As illustrated in Figure 1, the pipeline consists of three sequential stages: Dynamic Information Tree Construction, Automated Task Generation, and Adaptive Evolvement Loop. Detailed prompts and configuration settings are provided in Appendix A.

### 3.1 Dynamic Information Tree Construction

To ensure that the evaluation reflects the complexity and timeliness of the real internet, DR-Arena directly scrapes and structures real-world websites into a verifiable knowledge base, as illustrated by the instantiated example in Figure 2.

To ensure task diversity, the Examiner first samples a seed topic from Google Trends[1] and navigates to a specific sub-topic, e.g. *"Handheld Consoles"* $\rightarrow$ *"Nintendo History"*. Then, it generates a relevant search query to retrieve top websites from the open web. Among the top returned URLs, the Examiner selects a high-quality informative website to use as a "Root Node" ($v_{root}$) for the Information Tree, which is typically a comprehensive guide or hub page with rich outbound links

From $v_{root}$, the system constructs an Information Tree by scraping this page and its linked neighbors. We could represent the Information Tree as a directed graph $G = (V, E)$. Each node $v_i \in V$ represents a real webpage, containing the full text content $C_i$ and metadata (URL, title). An edge $e_{ij} \in E$ exists if node $v_i$ contains a hyperlink to $v_j$. To capture inter-node relationships $R_{ij}$, we enrich edges with semantic context by analyzing anchor text and surrounding content to derive relationship types such as varieties or processes.

While we limit the Information Tree size at first, the tree can be expanded in two orthogonal directions during the process: Depth Expansion recursively crawls specific branches to depth $d$ to establish reasoning chains (e.g., Handheld Console History $\rightarrow$ Specific Console $\rightarrow$ Technical Specs) supporting *Multi-hop Reasoning Tasks*; Width Expansion captures sibling nodes sharing the same parent and semantic relationship (e.g., multiple handheld consoles listed on the history page) to support aggregation tasks.

### 3.2 Automated Task Generation

Leveraging the Information Tree, the Examiner dynamically generates questions to probe two core capabilities of DR agents: multi-hop reasoning (*depth*) and information gathering (*width*).

**Depth and Width Principles.** When generating questions, we utilize the definition of *depth* and *width* from Lan et al. (2025). First, depth refers to the search steps required, which could be directly visualized by the tree depth. For example, deep reasoning tasks mask the target entity's name, requiring multi-hop deduction through relationship chains (e.g., "Identify the 1989 handheld console designed by the creator of the Game & Watch series"). Second, width refers to the number of information units to be searched, which can be visualized as tree width. Wide coverage tasks require synthesizing attributes across sibling nodes, demanding comprehensive aggregation through parallel entity comparisons (e.g., "Compare the battery specifications of the identified console against its market competitors"). Collectively, *depth* and *width* constitute two fundamental capabilities of DR agents: reasoning and information gathering.

**Task Generation.** The Examiner then generates a challenging question based on the Information Tree. First, it randomly selects a node from the tree and performs a validation check: if the current node lacks sufficient depth (no children) or width (no siblings) to support a challenging query, the system automatically triggers the crawler to

[1] https://github.com/pat310/google-trends-api/wiki/Google-Trends-Categories

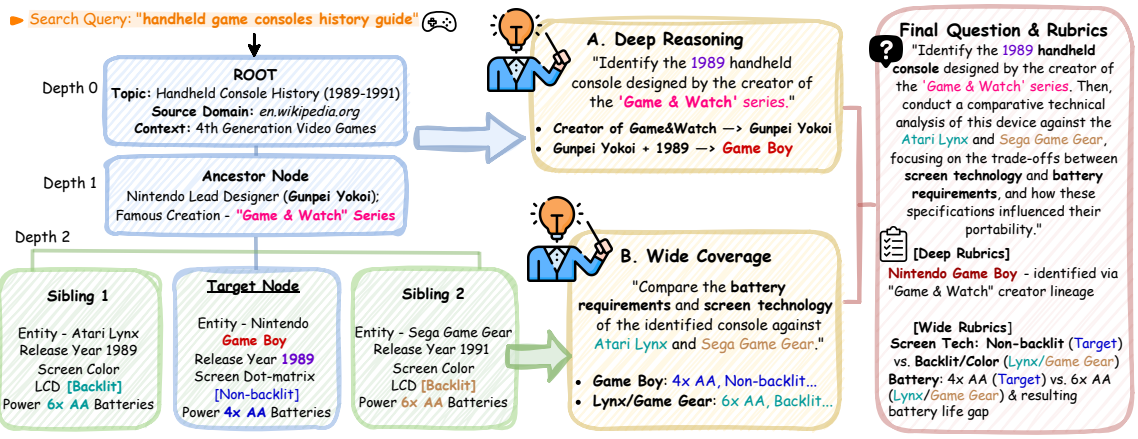


Figure 2: Automated Task Generation via Dynamic Information Trees.

fetch fresh data, expanding the tree structure on-the-fly. Then, the Examiner drafts a challenging query to probe both depth and width capabilities. To prevent candidates from taking shortcuts, the Examiner employs a de-contextualization strategy: When generating queries, the Examiner is strictly prohibited from mentioning specific filenames or website titles.

Since the ground truth ranging from logical lineages to attribute values is structurally encoded in the tree, the Examiner simultaneously synthesizes a list of rubrics that is factually correct and traceable to source URLs. This list includes a *Checklist-Depth* for verifying logical identity and a *Checklist-Width* for assessing data completeness. Representative task examples are provided in Appendix B.

### 3.3 Adaptive Evolvement Loop

After task generation, the Examiner assesses the candidate responses against the rubrics and enters the Adaptive Evolvement Loop, where it dynamically decides whether to draft follow-up questions to push candidates to their capability boundaries and hence make the performance gaps more visible.

| Failure Tag | Diagnostic Criteria |
|---|---|
| DEPTH | **Logic Failure.** Failed to identify the correct core entity due to a broken reasoning chain. |
| WIDTH | **Coverage Failure.** Failed to aggregate specific attribute details (Data Gap). |
| BOTH | **Systemic Failure.** Failed on both logical identification and factual completeness. |
| NONE | **Soft Gap.** The loss was determined solely by soft filters (formatting or utility preferences). |

Table 1: **Taxonomy of Failure Types.** The Examiner provides a diagnostic tag in verdicts.

**Evidence-Based Judgment.** Evaluating open-ended reports is inherently challenging due to the lack of ground-truth data. DR-Arena addresses this by employing the Examiner as a Judge, utilizing the generated rubrics via a strict two-stage protocol: First, the Examiner verifies *Hard Constraints* by checking against the list of rubrics. A critical error, such as misidentifying the core entity (logic error) or omitting mandatory data points (coverage gap), results in immediate penalties. Then, the Examiner evaluates *Soft Constraints*, comparing user experience aspects such as presentation quality, formatting, information density, and helpfulness.

Instead of a simple binary win/loss, we adopt a tiered adjudication system to capture the magnitude of performance gaps with finer granularity. When a candidate (A or B) wins, the verdict distinguishes whether it is a decisive win or a marginal win. Ties are also classified into High Quality (mutual success) or Low Quality (mutual hallucination) to inform the subsequent evolution strategy.

$$\text{Verdict}(A,B) \in \begin{cases} \textsc{Much Better} \text{ (Decisive Win)} \\ \textsc{Better} \text{ (Marginal Win)} \\ \textsc{Tie}: \begin{cases} \text{Both High Quality} \\ \text{Both Low Quality} \end{cases} \end{cases} \quad (1)$$

Guided by the Depth and Width principles, the Examiner also performs a failure type analysis on the losing side's response: Whether it is a Depth failure, a Width failure, or both/none. Detailed explanations on these failure types are outlined in Table 1. This distinction is critical as it disentangles reasoning deficits from retrieval gaps, providing the necessary signal for the Evolvement Loop to intelligently target the specific weakness in the subsequent rounds.

| Adjudication Verdict | Diagnostic Signal | Evolution Action | Strategic Rationale |
|---|---|---|---|
| **Tie** (High Quality) | N/A | **Pressure Test** ($D \uparrow 1$ & $W \uparrow 1$) | Current task too easy; find ceiling. |
| **Tie** (Low Quality) | N/A | **Backtrack** (Move to Parent) | Current task too hard; re-establish baseline. |
| **Winner Decided** | DEPTH (Logic Failure) | **Probe Depth** ($D \uparrow 1$) | Challenge loser's reasoning capabilities. |
| | WIDTH (Coverage Failure) | **Probe Width** ($W \uparrow 1$) | Challenge loser's information coverage. |
| | BOTH / NONE | **Pressure Test** ($D \uparrow 1$ & $W \uparrow 1$) | Ambiguous failure; increase difficulty. |

Table 2: **The Evolvement Loop Transition Matrix.** The system dynamically adjusts task complexity (Depth $D$ and Width $W$) based on adjudication verdict and diagnostic failure type of the losing agent.

**Follow-up Rounds.** After each round, the Examiner analyzes the verdict and failure type to operate on a targeted probing strategy to accelerate performance divergence and reach a decisive verdict, as outlined in Table 2, if a High-quality Tie occurs, the system triggers a pressure test, increasing both depth ($D$) and width ($W$) parameters to locate the capability ceiling. Conversely, if a marginal win occurs, the system aggressively targets the loser's specific weakness, probing either depth or width.

This adaptive mechanism ensures the system efficiently converges to a verdict by continuously pushing agents toward their specific breakdown points. The loop terminates when a decisive "Much Better" verdict is reached, the cumulative score difference exceeds a threshold, or the maximum round limit is met. Complete pseudo-code and a full match walkthrough are in Appendix C and D.

## 4 Experiments

### 4.1 Experimental Setup

**Model Choice.** From the pool of 13 models listed on LMSYS Search Arena (as of Dec 3, 2025), we select 6 representative models covering the latest or best-performing versions from major model families: GPT-5.1-Search (OpenAI), Gemini-2.5-Pro-Grounding (Google), o3-Search (OpenAI), Grok-4-Search (xAI), Perplexity-Sonar-Pro-High (Perplexity), and Claude-Opus-4.1-Search (Anthropic). As good-performing LLMs tend to perform better in LLM-as-a-judge tasks (Zheng et al., 2023), we choose Gemini-3-Pro as the fixed Examiner, which is currently the best model on LMSYS *Text* Arena.

**Baselines.** We compare DR-Arena against six recent benchmarks for evaluating DR search agents, including BrowseComp (Wei et al., 2025), DeepResearch Bench (Du et al., 2025), LiveNewsBench (Zhang et al., 2026), LiveSearchBench (Zhou et al., 2025), LiveResearchBench (Wang et al., 2025), and Deep Research Bench (FutureSearch et al., 2025). Details of these benchmarks are shown in Appendix E. As a reference, we utilize the LMSYS Search Arena scores as the human-annotated ground truth for evaluating alignment of the resulting rankings. They rely on large-scale blind human comparisons to establish trustworthy model rankings.

**Tournament Configuration.** To maximize computational efficiency, we conduct a Swiss-style tournament: For $n$ participants, instead of pairing each participant with $(n-1)$ others, a Swiss tournament pairs each player with $\lceil \log_2(n) \rceil$ players of similar rankings without repeats. This design effectively reduces computational costs of ranking $n$ models from $O(n^2)$ to $O(n \log_2(n))$. To mitigate bias, the tournament begins with a randomized initialization, followed by 4 rounds of dynamic pairing. While the minimum number of rounds recommended by the Swiss tournament design is $\lceil \log_2 6 \rceil = 3$, we deliberately oversample slightly to reduce variance from early-stage stochasticity. To ensure both domain robustness and fair comparison, each pairing competes across the same set of 30 pre-constructed Information Trees derived from diverse real-world websites (refer to Appendix F for detailed topic and domain distributions).

**Scoring Mechanism.** Following competitive gaming standards (Elo and Sloan, 1978), we utilize the Elo rating system updated via the Bradley-Terry (BT) model (Bradley and Terry, 1952). Elo scores are dynamically updated after each round to guide the subsequent pairings, ensuring that models continuously face opponents of comparable strength.

### 4.2 Main Results: Alignment with Humans

We evaluate the reliability of DR-Arena by benchmarking its derived rankings against the human-verified LMSYS Search Arena scores.

**Leaderboard Consistency.** Table 3 presents the comparative rankings derived from DR-Arena and LMSYS Search Arena. DR-Arena achieves approximate alignment with human judgment, recovering the exact hierarchy for most models. Figure 3 plots Elo scores derived from Search Arena against

| Model | Search Arena | | DR-Arena | |
|---|---|---|---|---|
| | Elo | Rank | Elo | Rank |
| GPT-5.1-Search | 1201 | **1** | 1084 | **1** |
| Gemini-2.5-Pro-Grounding | 1142 | **2** | 1054 | **2** |
| o3-Search | 1139 | **3** | 1041 | **3** |
| Grok-4-Search | 1138 | **4** | 958 | **4** |
| Claude-Opus-4.1-Search | 1130 | 5 | 921 | 6 (↓) |
| Perplexity-Sonar-Pro-High | 1125 | 6 | 942 | 5 (↑) |

Table 3: Comparison between DR-Arena Leaderboard and LMSYS Search Arena.

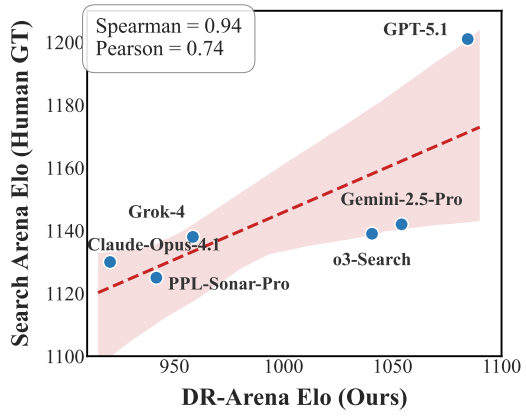


Figure 3: Leaderboard Correlation Analysis.

those derived from DR-Arena. We observe a linear mapping, confirming that our automated metrics effectively proxy human preference.

While the overall alignment is high, a minor rank swap occurs between Perplexity-Sonar-Pro (#5 in DR-Arena) and Claude-Opus-4.1 (#6). As detailed in Appendix G, this divergence could be an example of the *Factuality-Fluency Trade-off*: Our qualitative analysis reveals that the automated Examiner strictly penalizes hallucinated numbers that human annotators may overlook, being swayed by Claude's superior writing style.

**Superiority over Static Benchmarks.** We also calculate the correlation scores between LMSYS Search Arena and other benchmark datasets. Results are shown in Table 4. We observe that DR-Arena achieves the SOTA Spearman correlation of 0.94, and a Pearson correlation of 0.74, significantly outperforming static datasets like Deep Research Bench (FutureSearch) and LiveResearchBench, which show negative correlations. Notably, among all benchmarks, DR-Arena is the only one that requires zero human intervention, automating the ground-truth generation via Dynamic Information Trees to assess factual correctness on the live web. We hypothesize that this high alignment with humans stems from our *Dynamic Investigation* process, which simulates the iterative nature of human research, capturing nuances like error recovery and synthesis that static QA metrics could miss.

| Benchmark | N | Pearson | Spearman |
|---|---|---|---|
| **DR-Arena (Ours)** | **6** | **0.74** | **0.94** |
| BrowseComp (Wei et al., 2025) | 5 | 0.70 | 0.76 |
| DeepResearch Bench (Du et al., 2025) | 4 | 0.63 | 0.40 |
| LiveNewsBench (Zhang et al., 2026) | 4 | 0.46 | 0.20 |
| LiveSearchBench (Zhou et al., 2025) | 4 | -0.23 | -0.11 |
| LiveResearchBench (Wang et al., 2025) | 4 | -0.47 | -0.63 |
| Deep Research Bench (FutureSearch et al., 2025) | 6 | -0.86 | -0.90 |

Table 4: **Comparison with SOTA Benchmarks.** N denotes the number of models used for calculation.

## 4.3 Ablation Studies

To validate the design choices of DR-Arena, we conduct ablation studies focusing on three core components: tree-guided task generation, rubric-based judgments, and the evolvement loop.

| Configuration | Question Quality | Rubric Accuracy |
|---|---|---|
| **Full Tree (Ours)** | **89.0%** | **88.0%** |
| No Logic Chain | 2.0% | 8.0% |
| Flat Context | 9.0% | 4.0% |

Table 5: Human Preference Rates on Task Generation.

**Tree-guided Task Generation.** To verify the effectiveness of Information Trees, we conduct a blind human study comparing tree-based task generation against two ablation baselines: (1) Flat Context: Providing the Examiner with unstructured search snippets without topological edges. (2) No Logic Chain: Providing node content but withholding ancestor context (the reasoning path). For 50 sampled cases, two expert annotators performed a blind three-way selection to choose the generated question that best necessitated both Multi-hop Logic (Depth) and Multi-source Synthesis (Width) and the verification checklist that best captured the ground truth. Results are presented in Table 5. We observe that the full tree approach achieves the best performance, which has a selection rate of 89% for question generation and 88% for rubric construction. In contrast, the Flat Context baseline often degenerates into shallow factoid lookups, confirming that topological structure is essential for constructing valid deep research tasks.

**Rubric-Based Judgments.** We evaluate the adjudication mechanism by removing the generated Checklists and forcing the Examiner to judge based on internal knowledge and optional search. As

| Configuration | Pearson | Spearman |
|---|---|---|
| **DR-Arena (Full)** | **0.74** | **0.94** |
| *Impact of Adjudication* | | |
| w/o Rubric (Intuition Judge) | 0.72 | 0.83 |
| *Impact of Evolvement Loop* | | |
| w/o Evolvement (Round 1 only) | 0.41 | 0.77 |
| w/o Evolvement (Round 2 only) | 0.68 | 0.94 |

Table 6: **Ablation Results.** Removing the rubrics or the evolvement loop significantly degrades alignment.

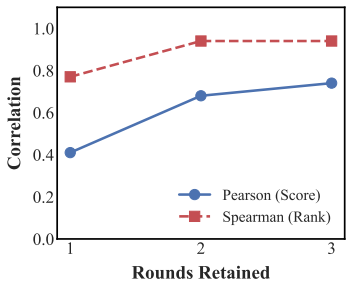


Figure 4: Evolution of Correlation across Rounds.

shown in Table 6, the Spearman correlation drops from 0.94 to 0.83. When examining the mismatched judgments closely, we see that, while the intuition-based judge can capture stylistic quality, it struggles to detect subtle hallucinations in synthesis. This demonstrates that Evidence-Based Adjudication is critical for reducing judge hallucination and ensuring rigorous alignment with human standards.

**The Evolvement Loop.** We assess the impact of the multi-round dynamic investigation by cutting off investigation rounds early and present the results in Table 6. The results derived solely from the First Round are insufficient, where Spearman correlation drops from 0.94 to 0.77, confirming that single-turn interactions fail to distinguish closely matched models, often resulting in ambiguous ties or noisy verdicts. By Round 2, the Spearman ranking correlation jumps to 0.94, demonstrating that the pressure test mechanisms are highly efficient at sorting agents into the correct hierarchy. The loop remains essential beyond this point, which we show with a more straightforward plot in Figure 4: The Pearson correlation, which measures the linearity of the score gaps, continues to improve from 0.68 to 0.74. This confirms that the subsequent rounds are important for calibrating the magnitude of the performance difference.

**Cross-Examiner Validation.** To further examine whether the reliability of DR-Arena depends on a specific Examiner model, we conduct a Cross-Examiner validation experiment. We randomly sample 50 matches from the tournament logs and re-adjudicate them using two alternative frontier LLMs, GPT-5.2-Chat and Claude-Opus-4.6. We then compare their verdicts against the original adjudications produced by Gemini-3-Pro.

| Judge Model | Percent Agreement | Cohen's Kappa |
|---|---|---|
| GPT-5.2-Chat | 93.02% | 0.901 |
| Claude-Opus-4.6 | 88.70% | 0.839 |

Table 7: **Cross-Examiner Validation.** Agreement rates between alternative Examiners and the original Examiner on 50 sampled matches.

| Component | Evaluation Criterion | Result |
|---|---|---|
| **Task Generation** | Question Validity (Structure) | 90.6% |
| | Rubric Factuality (Ground Truth) | 89.1% |
| **Examiner Process** | Verdict Alignment (Cohen's $\kappa$) | 0.91 |
| | Transition Logic Accuracy | 96.9% |
| | Stop Condition Efficiency | 92.2% |
| **Human Annotators** | Inter-Annotator Agreement (Cohen's $\kappa$) | 0.88 |

Table 8: **Human Audit Results.** Expert annotators ($N = 2$) evaluated the quality of automated components across 30 randomly sampled matches (64 turns).

As shown in Table 7, the alternative Examiners achieve high agreement with the original Examiner, with Percent Agreement rates of 93.02% and 88.70%, and Cohen's Kappa scores of 0.901 and 0.839, respectively. These results indicate that DR-Arena's judgments are highly consistent across diverse frontier LLMs, suggesting that the evaluation signal is not primarily driven by the stylistic bias of a particular judge model. We hypothesize that this robustness stems from the structured nature of our framework: unlike fully open-ended judging, the Information Tree provides grounded rubrics that function as an explicit answer key, substantially reducing variance in the adjudication process.

### 4.4 Validation via Human Study

To validate the reliability of our automated components, we conduct a rigorous human study where two expert annotators audited a random sample of 30 full-match logs (covering 64 interaction turns). Detailed human study setups and interface demos are provided in Appendix H.

During the study, we ask the annotators to evaluate the match quality while focusing on two aspects: task generation and Examiner judgment. The results are summarized in Table 8, which strongly validate the quality of both the generation and adjudi-

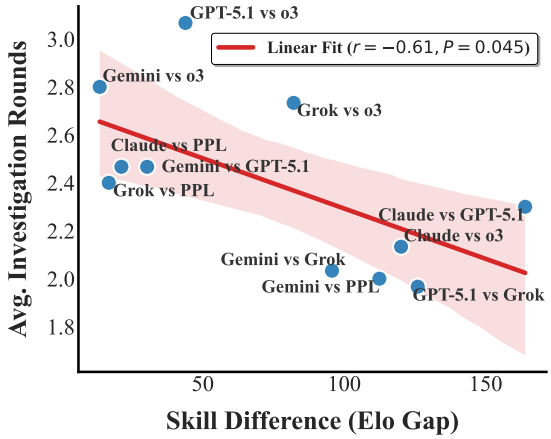


Figure 5: Relationship between Skill Gap and Rounds.

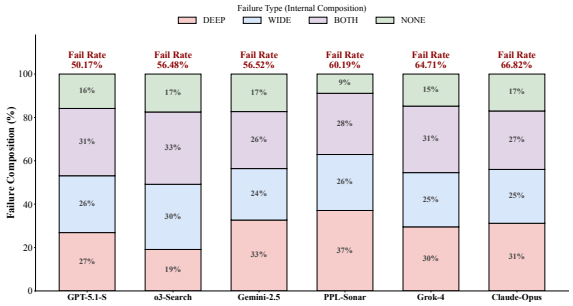


Figure 6: Per-Model Performance Profile. Models are ordered by their overall failure rates (at the top).

cation pipelines. First, regarding Task Generation, 90.6% of the generated questions are confirmed to strictly adhere to the Depth and Width principles and 89.1% of the rubrics are verified as factually correct based on the source URLs. Second, regarding the judgment process, the automated verdicts achieve substantial alignment with human experts' judgments, yielding a Cohen's Kappa agreement of 0.91. The two human annotators themselves achieve a Cohen's Kappa agreement of 0.88 on verdict selection, indicating high inter-annotator consistency. Furthermore, the adaptive mechanism is proved to be highly reliable: 96.9% of the Evolvement Loop transitions correctly follow the diagnostic logic, and 92.2% of the matches are deemed to have stopped at an efficient round count. These findings validate the automated Examiner as a trustworthy, scalable proxy for human adjudication.

## 5 In-depth Analysis

Beyond quantitative evaluation results, we also perform in-depth analysis to examine the efficiency of the dynamic loop and the error distribution profiles of different models. Extended statistical diagnostics, including the aggregate distribution of verdicts and the structural diversity of the generated Information Trees, are detailed in Appendix I.

### 5.1 Efficiency: The Skill Gap vs. Rounds

To examine the hypothesis of DR-Arena where follow-up rounds make performance gaps more apparent, we analyze the relationship between *skill difference* (Elo Gap) and *number of investigation rounds* in Figure 5. We quantify the linear relationship using the Pearson correlation coefficient ($r$) and assess statistical significance via the $p$-value. We observe a significant negative correlation ($r = -0.61$, $p = 0.045$) between average investigation rounds and skill gap, validating that pairs with smaller skill gaps indeed require more investigation rounds. In the graph, we see that the system converges rapidly for high-divergence pairs; for instance, matches involving *GPT-5.1-Search* vs. *Grok-4-Search* ($\Delta$Elo $\approx$ 126) average fewer than 2.0 rounds, as the stronger model quickly satisfies hard constraints while the weaker fails. Conversely, closely matched pairs such as *Gemini-2.5-Pro* vs. *o3-Search* ($\Delta$Elo $\approx$ 13.5) extend to 2.8 rounds. In these scenarios, the system automatically detects and triggers the Pressure Test mechanism, escalating complexity to force a fine-grained differentiation. Therefore, DR-Arena effectively functions as an efficient sorting algorithm, concentrating computational cost on the decision boundaries where it yields the highest information gain.

Notable outliers provide further insights into model architectures. The matchup between *GPT-5.1-Search* and *o3-Search* exhibits the highest investigation duration (3.07 rounds) despite a moderate Elo gap ($\Delta$Elo $\approx$ 44). We hypothesize this stems from Behavioral Homogeneity within the OpenAI family: Both models likely share similar data and reasoning answer patterns, which lead to more ties. The system then escalates difficulty with more rounds to uncover subtle differences in long-horizon synthesis capabilities, demonstrating that DR-Arena can distinguish even highly correlated models, albeit at higher computational cost.

### 5.2 Error Distribution: Depth vs. Width

To provide a closer look into each model's performance profile, we show a dissection of failure types in Figure 6. We define the Failure Rate as the ratio of non-winning rounds (including losses and ties) to the total number of rounds each model participated in. Ordering models by this metric reveals a clear robustness hierarchy, with *GPT-5.1-Search*

demonstrating the highest stability with a minimum failure rate of 50.17%.

Beyond aggregate rates, the normalized failure distribution exposes architectural trade-offs. *GPT-5.1-Search* exhibits symmetric errors (27% Depth, 26% Width), suggesting balanced capabilities without structural bottlenecks. In contrast, *o3-Search* shows asymmetry favoring logical deduction: despite higher total failures, it has the lowest DEPTH failure rate (19%), with limitations primarily in coverage breadth (WIDTH: 30%). Models like *PPL-Sonar-Pro* and *Grok-4* display the inverse pattern, with failures skewed toward DEEP reasoning deficits (37% and 30% respectively), indicating effective retrieval but inconsistent logical reasoning. The persistent BOTH failures (∼20-30%) across all models underscore the difficulty of simultaneously satisfying depth and width constraints. These profiles reveal capability trade-offs beyond aggregate metrics, enabling targeted model improvements.

## 6 Conclusion

We introduce DR-Arena, a fully automated framework for evaluating DR agents in dynamic environments. Existing static benchmarks suffer from temporal misalignment, penalizing agents that retrieve current information contradicting outdated ground truth. DR-Arena addresses this by constructing real-time Information Trees synchronized with the live internet. Beyond timeliness, it transforms evaluation into an active stress test through Adaptive Evolvement, efficiently isolating performance boundaries and disentangling failures in logical deduction (*Depth*) from deficits in information aggregation (*Width*). Achieving 0.94 Spearman correlation with human-verified leaderboards, DR-Arena offers a scalable, reliable alternative to costly manual evaluation for benchmarking next-generation autonomous research systems.

## Limitations

While DR-Arena provides a scalable alternative to human annotation, we acknowledge several limitations inherent to the framework.

First, although our cross-examiner validation shows strong agreement across different judge models, DR-Arena still relies on a single Examiner in the main evaluation pipeline, so judge-specific preferences cannot be fully ruled out.

Second, when the generated ground-truth rubrics conflict with the Examiner's internal parametric knowledge, there arises a risk that a judge could override the provided rubrics due to strong parametric priors. In some cases, the judge could also successfully disregard a mistake in the generated list, as shown in Appendix G.3. Overall, the judge's parametric knowledge conflicts remain a risk and future direction to explore.

Third, because DR-Arena operates over the live web and depends on commercial search APIs, exact reproduction may be affected by temporal changes in indexing, ranking, geoblocking, and webpage availability. This creates a trade-off between ecological realism and fully controlled comparison: while we share the same fixed Information Trees across agents for fair evaluation, source prioritization and search ranking may still disadvantage agents that rely on different but potentially more reliable sources. We reduce this risk by biasing tree construction toward high-reliability seed domains, while noting that fully freezing the search environment would re-introduce contamination risks.

Finally, although our evidence-based rubric correlates well with human preference on fact-heavy research tasks, it may undervalue creative synthesis or lateral thinking that deviates from the strict logical path of the generated tree.

## Acknowledgments

This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016). This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-01-001).

## References

Roshita Bhonsle, Rishav Dutta, Sneha Vavilapalli, Harsh Seth, Abubakarr Jaye, Yapei Chang, Mukund Rungta, Emmanuel Aboah Boateng, Sadid Hasan, Ehi Nosakhare, and Soundar Srinivasan. 2025. Auto-eval judge: Towards a general agentic framework for task completion evaluation. *Preprint*, arXiv:2508.05508.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345.

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

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Arpad E Elo and Sam Sloan. 1978. *The Rating of Chessplayers: Past and Present*. Arco Publishing, New York.

FutureSearch, Nikos I. Bosse, Jon Evans, Robert G. Gambee, Daniel Hnyk, Peter Mühlbacher, Lawrence Phillips, Dan Schwarz, and Jack Wildman. 2025. Deep research bench: Evaluating ai web research agents. *Preprint*, arXiv:2506.06287.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, and 7 others. 2025. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *Preprint*, arXiv:2506.21506.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: what's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

Tian Lan, Bin Zhu, Qianghuai Jia, Junyang Ren, Haijun Li, Longyue Wang, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025. Deepwidesearch: Benchmarking depth and width in agentic information seeking. *Preprint*, arXiv:2510.20168.

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See-Kiong Ng, and Tat-Seng Chua. 2025. FACT-AUDIT: an adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 360–381. Association for Computational Linguistics.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. 2026. Search arena: Analyzing search-augmented llms. *Preprint*, arXiv:2506.05334.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

OpenAI. 2025. Deep Research: Accomplish complex tasks with agentic analysis. https://openai.com/index/deep-research.

Perplexity AI. 2025. Perplexity AI: Answer engine for complex queries. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Webcpm: Interactive web search for chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8968–8988. Association for Computational Linguistics.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2025. A comprehensive survey of contamination detection methods in large language models. *Preprint*, arXiv:2404.00699.

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

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720. Association for Computational Linguistics.

Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, Wenlong Zhang, Philip Torr, and Dongzhan Zhou. 2026. Deep research arena: The first exam of llms' research abilities via seminar-grounded tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33341–33349.

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

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia

Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *Preprint*, arXiv:2504.12516.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha V. Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and Ke Wang. 2025. Widesearch: Benchmarking agentic broad info-seeking. *Preprint*, arXiv:2508.07999.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2026. Livenewsbench: Evaluating llm web search capabilities with freshly curated news. *Preprint*, arXiv:2602.13543.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. 2025. Auto-arena: Automating LLM evaluations with agent peer battles and committee discussions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4440–4463. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Heng Zhou, Ao Yu, Yuchen Fan, Jianing Shi, Li Kang, Hejia Geng, Yongting Zhang, Yutao Fan, Yuhao Wu, Tiancheng He, Yiran Qin, Lei Bai, and Zhenfei Yin. 2025. Livesearchbench: An automatically constructed benchmark for retrieval and reasoning over dynamic knowledge. *Preprint*, arXiv:2511.01409.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.

# A Prompts and Configuration Settings

To ensure reproducibility, we provide all prompt templates used in DR-Arena, including Automated Task Generation and Evidence-Based Judgment. In the templates below, dynamic content injected by the system at runtime is denoted by {{Blue Placeholders}}.

## A.1 Automated Task Generation

The Examiner uses the following prompt to transform a raw Information Tree into a structured "Depth & Width" research task. The system dynamically calculates word count constraints based on the tree depth and width.

**System Prompt: Task Generation**

```
# TASK: Generate a ''Depth & Width'' Search
    Evaluation Query

You are an expert at creating complex,
    multi-hop search queries designed to
    test the limits of Search Agents. Your
    goal is to synthesize a question that
    requires Logical Reasoning
    (Depth) to identify the subjects and
    Broad Information Aggregation
    (Width) to answer fully.

--- 1. THE HIDDEN KNOWLEDGE (Source
    Material) ---
*Note: This content is hidden from the test
    taker. It is only for you to formulate
    the question and the grading criteria.*

OVERALL DOMAIN/TOPIC: ''{{Root Topic}}''
A. Reasoning Chain (Background/Context):
{{Serialized Reasoning Nodes (Ancestors)}}
B. Target Answers (The Facts to Retrieve):
{{Serialized Target Nodes (Siblings)}}

--- 2. QUESTION GENERATION STEPS (READ
    CAREFULLY) ---
Rule 1: ABSOLUTE GROUNDING - CRITICAL
- YOU MUST generate the question based ONLY
    on the specific entities and facts
    found in the [Hidden Knowledge] above.
- STRICT PROHIBITION: Do NOT ignore the
    provided text.
- Relevance: The question MUST be relevant
    to the Overall Domain/Topic (''{{Root
    Topic}}''). Do not hallucinate
    unrelated topics.

Rule 2: COMPLETE DE-CONTEXTUALIZATION (No
    Leaking)
- FORBIDDEN: You MUST NOT mention the
    specific filename, website title,
    directory name, or document header
    found in the source.
- REQUIRED: Treat the provided text as just
    *one instance* of a universal fact. Ask
    about the *entities themselves*, not
    about the *document* describing them.
- Litmus
    Test: If the user needs the specific
    JSON file you read to understand the
    question, YOU FAILED. The question must
    be solvable using Google/Bing to find
    the *original primary sources*.

STEP 1: Deep Reasoning (The Filter)
- Analyze the [Reasoning Chain] to identify
    the specific logic, condition, or
    category that groups the target
    entities together.
- RULE: Do NOT mention the specific names
    of the [Target Entities] in the
    question.
- RULE: Use the [Reasoning Chain] logic to
    strictly define the group.

STEP 2: Wide Aggregation (The Scope)
- If the [Target Answers] contain multiple
    entities, the question MUST require
    reading and comparing information from
    ALL of them.
- The answer must not be resolvable by
    finding a single document; it must
    require aggregating details across all
    identified targets.

STEP 3: Synthesis (The Depth & Width
    Question)
- Combine Step 1 and Step 2 into a single,
    cohesive natural language question.
- CRITICAL: Ensure the question targets
    Publicly Verifiable
    Facts. Do not ask about obscure details
    that exist *only* within the specific
    phrasing of the provided source text.
    The question must be answerable by
    searching external, general web sources.

--- 3. CHECKLIST DEFINITIONS (CRITICAL) ---
STEP 1 Draft the Gold Standard
    Answer: Formulate a complete answer
    based on the [Hidden Knowledge].
STEP 2 Extract
    Checklists: Deconstruct the answer into
    specific verification points.
    - Checklist Width (Completeness &
    Details): Content: The Specific
    Attributes/Facts requested in the
    query. Purpose: Once the entity is
    found, did the agent gather *all* the
    requested scattered details?
    - Checklist Depth (Identity &
    Logic): Content: The Correct Entity
    Names + The Logic Validation. Purpose:
    Did the agent use the reasoning chain
    to find the *correct* person/thing?

--- 4. OUTPUT FORMAT (JSON) ---
Return the result in the following JSON
    format:
{
    ''question'': ''The final Depth & Width
    search query'',
    ''word_limit_instruction'': ''{{Dynamic
    Constraint String}}'',
```

```
    ''checklist_width'': [
        ''Specific Detail A for Entity 1'',
        ''Specific Detail B for Entity 1'',
        ...
    ],
    ''checklist_depth'': [
        ''Target Entity 1 Name + Logic
    Proof'',
        ...
    ],
    ''rationale'': ''Briefly explain how
    the question uses logic to mask
    entities (Depth) and requests scattered
    info (Width).''
}
```

## A.2 Evidence-Based Judgment

The Examiner acts as a Judge to evaluate two DR agents responses using the generated Rubrics.

**System Prompt: Judgment**

```
### Role: Super-User Evaluator (Simulating
    Human Preference)
Compare Response A and Response B to
    identify which search agent provides a
    better USER EXPERIENCE.
While accuracy is paramount, you must also
    heavily weigh comprehensiveness,
    formatting, and
    helpfulness -- traits that human users
    value in search engines like
    Perplexity, Gemini, or SearchGPT.

--- 1. QUERY & CONSTRAINT ---
Query: {{Question}}
Constraint: Maximum {{Word Limit}}
    words. (Note: Do not penalize slightly
    going over if the quality is high. Only
    penalize extreme verbosity).

--- 2. GROUND TRUTH CHECKLIST ---
[WIDTH-Completeness]: {{Checklist Width
    JSON}}
[DEPTH-Logic]: {{Checklist Depth JSON}}

--- 3. RESPONSES ---
=== Agent A ===
(Citation Count: {{Count A}})
{{Final Answer A}}
=== Agent B ===
(Citation Count: {{Count B}})
{{Final Answer B}}

--- 4. EVALUATION CRITERIA (Aligned with
    Human Preference) ---
Dimension 1: Accuracy (The Foundation)
- Core Entity Check: Determine if each
    agent passes the DEPTH Logic (Found the
    right entity?). (If BOTH fail this,
    it's a LOW TIE).
- Sub-Point Accuracy: Did the agent answer
    *all* parts of the prompt correctly?
    Determine if each agent passes the
    WIDTH Aggregation (Found the specific
    details?).
- If BOTH agents have significant
    hallucinations (even on different
    parts), consider a Low Quality
    Tie.

Dimension 2: User Utility & Completeness
    (The Experience)
- Helpfulness: Is the answer easy to read?
    Does it actually solve the user's
    underlying intent?
- Information
    Density: Unlike simple chatbots, Search
    Agents should provide rich
    context.
- Helpful Recovery: If the exact answer
    isn't in the context, did the agent try
    to synthesize *related* useful info?
- Citation Density: A higher citation count
    is generally preferred as it indicates
    better groundedness.

Dimension 3: Presentation & Structure
- Markdown Mastery: REWARD the use of Bold
    headers, Bullet points, and Tables.
- Scannability \& Directness: Can a user
    find the specific answer in 2 seconds?
    (BLUF - Bottom Line Up Front)?

--- 4. SCORING RUBRIC ---
- [[A/B_MUCH_BETTER]] (+2):
    - The winner found the correct Entity
    AND answered sub-points correctly (No
    Hallucinations).
    - The loser failed the Depth Logic
    (Wrong Entity) or missed major
    Checklists.
    - *Note: Do not give MUCH_BETTER if the
    winner has a factual error in a
    sub-point.*
- [[A/B_BETTER]] (+1):
    If winner has errors, cap at BETTER.
    - The "Flawed
    Winner": The winner got the Main Entity
    right, but missed a detail or
    hallucinated on a minor sub-point. The
    loser failed the Main Entity.
    - The "Style
    Winner": Both are factually accurate,
    but one has significantly better
    formatting/comprehensiveness.
    - The "Nuance Winner": Both failed
    slightly, but the winner's failure was
    less catastrophic than the loser's.
- [[Tie]]:
    - *High Quality*: Both gave perfect,
    well-formatted, accurate answers.
    - *Low Quality*: Both failed to find
    the core entity or both hallucinated
    significantly.

Error Diagnosis
- If there is a loser, identify WHY they
    lost.
- DEPTH: Failed logic/identity (Wrong
    Entity).
- WIDTH: Failed detail aggregation (Missing
    Facts).
```

```
- BOTH: Failed both depth logic and width
    details.
- NONE: No hard checklist failures, when
    the winner won solely on Soft Filters
    like citations/formatting.

--- 5. OUTPUT FORMAT (JSON) ---
{
    ''verdict'': ''[[A_MUCH_BETTER]]'' OR
    ''[[A_BETTER]]'' OR ''[[Tie]]'' ... ,
    ''tie_quality'': ''HIGH'' OR ''LOW'' OR
    ''N/A'',
    ''loser_failure_type'': ''DEPTH'' OR
    ''WIDTH'' OR ''BOTH'' OR ''NONE'',
    ''reasoning'': ''First, verify Depth
    Logic for both. Then, compare
    Width/Completeness...''
}
```

## B Example Generated Tasks

This section presents five diverse examples of "Depth & Width" research tasks generated by the Examiner. These examples demonstrate the system's ability to construct complex queries requiring multi-hop reasoning (Depth) and broad information aggregation (Width) across different domains.

**Example 1: Local Business Structure Analysis**

**[QUESTION]**
Locate the gymnastics organization in the New York/New Jersey area that structures its competitive teams into three specific tiers: an in-house 'Club Team' that competes exclusively against the organization's other branches, a 'USA-IGC' program capped at 3 training days per week to accommodate other sports, and a third high-performance program that explicitly prohibits athletes from participating in other activities. Provide the name of this organization, the specific name of the high-performance program, and a list of all the cities where they currently operate gyms.

---

**[CHECKLIST / GRADING KEY]**
**Entities to find**: Target Entity: **Gold Medal Gymnastics** (or GMGC / Gold Medal Gymnastics & Ninja).
**Logic Proof**: Identified via the unique combination of an internal-only 'Club Team' and a 'Junior Olympic' program that prohibits other sports.
**Specific Details**:
- High-Performance Program Name: **Junior Olympic Program** (or Junior Olympic Team)
- City 1: Centereach
- City 2: Garden City
- City 3: Huntington
- City 4: Levittown
- City 5: Rocky Point
- City 6: Short Hills
- City 7: Smithtown

**Example 2: Cross-Domain Synthesis (Marketing & Labor Stats)**

**[QUESTION]**
Locate the analysis by 'Marketing Eye Atlanta' regarding a viral charity campaign that involved a 'bucket of iced water' and the participation of 'the most feared woman in fashion.' Identify the campaign and list the five specific elements cited that made it go viral. Then, identifying the specific occupation defined by the Bureau of Labor Statistics (BLS) as professionals who 'create and maintain a positive public image for the clients they represent,' provide the following data points from the 2024 occupational profile: the 2024 median annual pay, the job outlook percentage for the 2024–34 decade, and the projected numeric employment change for that same period.

---

**[CHECKLIST / GRADING KEY]**
**Entities to find**: Target Entity 1: **ALS Ice Bucket Challenge** (identified via Marketing Eye Atlanta & Anna Wintour clue), Target Entity 2: **Public Relations Specialists** (identified via job description match).
**Specific Details**:
- Element 1: Emotional AND logical appeal
- Element 2: Celebrity certification
- Element 3: Amusement

- Element 4: Simplicity
- Element 5: Shareability
- 2024 Median Annual Pay: **$69,780**
- Job Outlook (2024–34): **5%** (Faster than average)
- Employment Change (2024–34): **15,000**

### Example 3: Tech Policy & International Specifications

**[QUESTION]**
Analyze the language support specifications for the 'Live Translation' feature powered by Apple Intelligence as detailed in the user guide covering macOS Tahoe and iOS 26. Determine which three languages are explicitly compatible with Live Translation in the Messages app but are not listed as supported for Live Translation in Phone and FaceTime calls. Following this, consult the terms and conditions for the Apple Store Singapore to find the specific return policy for products purchased in volume (defined as an aggregate of more than four items). Provide the names of the three distinct languages, the number of days allowed for a volume return, and the restocking fee percentage charged.

---

**[CHECKLIST / GRADING KEY]**
**Entities to find**: Target Entity: **Apple Intelligence Live Translation** (Logic: Compare Messages support list vs. Phone/FaceTime support list), Target Entity: **Apple Store Singapore Policy** (Logic: Identify specific terms for 'returned in volume' > 4 items).
**Specific Details**:
- Language 1: **Dutch**
- Language 2: **Turkish**
- Language 3: **Vietnamese**
- Volume Return Deadline: **7 days**
- Volume Restocking Fee: **25%**

### Example 4: Medical Service Scope & Pricing

**[QUESTION]**
Investigate the remote medical second opinion service operated by the joint venture between Cleveland Clinic and Amwell. Provide a detailed breakdown of its geographic scope by listing the U.S. states eligible for the full 'Concierge Plus' virtual visit, those restricted to the written report only, and the specific states where the service is unavailable. Additionally, report the service costs for U.S. patients versus international patients, and enumerate the specific countries where the international service is prohibited.

---

**[CHECKLIST / GRADING KEY]**
**Entities to find**: Target Entity: **Virtual Second Opinions (VSO)** by 'The Clinic' (Cleveland Clinic & Amwell Joint Venture).
**Specific Details**:
- Identifies 'Concierge Plus' (Virtual Visit) states: Ariz., Calif., Colo., Conn., Fla., Ga., Ill., Ind., Ky., Mich., N.C., N.J., N.Y., Ohio, Pa., S.C., Tenn., Texas, Va., Wisc., W.V.
- Identifies 'Written Report Only' states: Alaska, Ala., Ark., D.C., Del., Hawaii, Iowa, Idaho, Kan., La., Mass., Md., Minn., Mo., Miss., Mont., N.D., Neb., N.H., N.M., Nev., Okla., Ore., Utah, Vt., Wash., Wyo.
- Identifies Excluded U.S. states: **Maine, Rhode Island (R.I.), South Dakota (S.D.)**
- States U.S. Pricing: **$1,690** (Report only) and **$1,990** (Report + Virtual Visit)
- States International Pricing: **$4,500 USD**
- Lists Excluded Countries: Australia, China, Germany, Denmark, Greece, Iran, North Korea, South Korea, Kazakhstan, Malaysia, Russian Federation, Sweden, Turkey

### Example 5: Cultural Analysis (Music History)

**[QUESTION]**
In a 1988 review by Steve Pond (RS 537), two distinct Los Angeles-bred acts were compared: one described as an "unprolific" artist making "immaculate pop music" with lushness borrowed from soundtracks, and the other a "young and restless" band labeled the "true heir to Led Zeppelin" but stripped of "fairy-tale whimsy." Despite their differences, the reviewer noted that both artists' respective albums from that year were populated by "recognizable, real people." Identify these two acts and their corresponding albums. Then, based on the specific descriptions in the review, list the following tracks: for the band, the two songs characterized as "hard-boiled riff rockers" and the acoustic song deemed a "worthy Left Coast successor to Walk on the Wild Side"; for the songwriter, the two songs presenting "bucolic views of a childhood in New Orleans," the two "naive, devoted love songs," and the final track described as a "chilling, coldblooded moment" involving a message to his son.

---

**[CHECKLIST / GRADING KEY]**
**Entities to find**: Band: **Jane's Addiction** (Album: Nothing's Shocking), Songwriter: **Randy Newman** (Album: Land of Dreams). **Logic Proof**: Validates the "heir to Led Zeppelin" vs "immaculate pop" comparison from the 1988 Steve Pond review.
**Specific Details**:
- Band Song (Hard-boiled): "Had a Dad"
- Band Song (Hard-boiled): "Standing in the Shower... Thinking"
- Band Song (Successor): "Jane Says"
- Songwriter Song (Bucolic): "Dixie Flyer"
- Songwriter Song (Bucolic): "New Orleans Wins the War"
- Songwriter Song (Love): "Falling in Love"
- Songwriter Song (Love): "Something Special"
- Songwriter Song (Chilling): "I Want You to Hurt Like I Do"

## C Algorithm of Evolvement Loop

The core logic of the DR-Arena Dynamic Investigation is detailed in Algorithm 1. The system manages the state of the Information Tree $T$, the current path $P$ (representing Depth), and the width constraint $W$.

**Algorithm 1** DR-Arena Adaptive Evolvement Loop

**Require:** Information Tree $T$, Agents $M_A, M_B$, Examiner $E$
**Ensure:** Final Verdict $V$ and Scores $S_A, S_B$
**Initialize:** Path $P \leftarrow$ RandomStart($T$), Width $W \leftarrow 2$
**Initialize:** Scores $S_A \leftarrow 0, S_B \leftarrow 0$
**while** $|S_A - S_B| <$ Threshold **and** Rounds $<$ Max **do**
  *// Phase 1: Environment Check & Expansion*
  **if** $|\text{Siblings}(P)| < W$ **then**
    $T \leftarrow$ Crawler.ExpandWidth($P$, target $= W$)
  **end if**
  *// Phase 2: Context Extraction & Task Generation*
  $C_{depth} \leftarrow$ Ancestors($P$)
  $C_{width} \leftarrow$ Siblings($P$, limit $= W$)
  $Q, \mathcal{R} \leftarrow E$.Generate($C_{depth}, C_{width}$) ▷ Gen. Question & Rubric
  *// Phase 3: Agent Execution & Adjudication*
  $Traj_A \leftarrow M_A(Q)$; $Traj_B \leftarrow M_B(Q)$
  $V, F_{type} \leftarrow E$.Judge($Traj_A, Traj_B, \mathcal{R}$)
  Update $S_A, S_B$ based on $V$
  *// Phase 4: Adaptive State Transition*
  **if** $V$ is TIE **then**
    **if** $V$ is LOW_QUALITY **then** ▷ Both Hallucinated
      $P \leftarrow$ Parent($P$) ▷ **Backtrack**
      $W \leftarrow \max(2, W - 1)$
    **else** ▷ High Quality Tie
      $W \leftarrow W + 1$ ▷ **Pressure Test**
      ATTEMPTDESCEND
    **end if**
  **else** ▷ Winner Decided
    **if** $F_{type}$ is DEPTH **then** ▷ Logic Failure
      ATTEMPTDESCEND
    **else if** $F_{type}$ is WIDTH **then** ▷ Coverage Failure
      $W \leftarrow W + 1$
    **else** ▷ Ambiguous Failure
      $W \leftarrow W + 1$; ATTEMPTDESCEND
    **end if**
  **end if**
**end while**

**procedure** ATTEMPTDESCEND
  **if** $P$ is Leaf Node **then**
    $T \leftarrow$ Crawler.ExpandDepth($P$)
  **end if**
  **if** $P$ has Children **then**
    $P \leftarrow P +$ RandomChild($P$)
  **end if**
**end procedure**

## D Full Match Walkthrough

To illustrate the dynamic nature of the DR-Arena, we present a complete 5-round match trace between Agent A (Perplexity-Sonar-Pro) and Agent B (Claude-Opus-4.1).

This match, centered on the topic of "Handheld Game Consoles," demonstrates the *Adaptive Evolvement* mechanism: as agents succeed, the system increases complexity (Depth/Width); when they fail, the system targets their specific weaknesses. The match concludes via the *Mercy Rule* when Agent B establishes a decisive lead.

**Match Trace: Handheld Game Consoles (Rounds 1-5)**

**Match Metadata**
**Topic**: Handheld game console
**Agents**: Agent A (Perplexity-Sonar-Pro) vs. Agent B (Claude-Opus-4.1-Search)
**Final Result**: Agent B Wins (Score 5.0 vs 2.0)

**=== ROUND 1 (Initialization) ===**
**State**: Depth 2 | Width 2
**Question**: Identify the two handheld game consoles released in the 1990s that were uniquely designed to play the exact same physical game media as their manufacturer's home console counterparts. Provide launch price, battery count, and battery life.

**Agent A Response**: Correctly identified **NEC TurboExpress** and **Sega Nomad**. Used bullet points for Price ($249.99/$179.99), Batteries (6 AA), and Life ( 3 hrs).
**Agent B Response**: Correctly identified the same entities and data, but used narrative paragraphs.
**Verdict**: [[A_BETTER]]
**Reasoning**: "Agent A is the winner based on Dimension 3 (Presentation). By using bullet points... Agent A made the answer immediately scannable... Agent B provided the same info but buried it within narrative paragraphs."
**Evolution**: Winner is A → Action: **Pressure Test (Depth+1 & Width+1)**

**=== ROUND 2 (Complexity Increase) ===**
**State**: Depth 2 | Width 3
**Question**: Identify the 8-bit handheld (1989) that succeeded 'Game & Watch' and included *Tetris*. Contrast battery specs against Atari Lynx and Sega Game Gear.

**Agent A Response**: Identified Game Boy. Estimated battery life as "10-15 hours" based on early specs. Used placeholders for citations.
**Agent B Response**: Identified Game Boy. Estimated battery life as "15-35 hours" (closer to retrospective consensus). Provided valid URL citations.
**Verdict**: [[B_BETTER]]
**Reasoning**: "Agent B wins based on Content Accuracy. The prompt asked for 'retrospective comparisons'... Agent B correctly captures the legendary efficiency (30+ hours) vs competitors... Agent A provided a conservative estimate."
**Evolution**: Winner is B → Loser Failure: WIDE → Action: **Wide+1 (Increase Context Width)**

**=== ROUND 3 (Technical Detail Test) ===**
**State**: Depth 2 | Width 4

**Question**: Identify three commercial color handhelds (1989-1991). Provide Price, CPU Architecture, and Battery Count.

**Agent A Response**: Identified Lynx, Game Gear, TurboExpress. Correctly identified Lynx CPU as **8-bit (6502 derivative)**. Excellent formatting.
**Agent B Response**: Identified same units. Incorrectly stated Lynx CPU was **"16-bit 65SC02"**.
**Verdict**: [[A_BETTER]]
**Reasoning**: "Agent A wins on Technical Accuracy... Agent A correctly identified the Atari Lynx CPU as an 8-bit 6502 derivative. Agent B incorrectly stated it was a 16-bit processor (a common misconception)."
**Evolution**: Winner is A → Action: **Pressure Test**

---

**=== ROUND 4 (Differentiation Test) ===**
**State**: Depth 2 | Width 5
**Question**: Identify the Game Boy's two primary competitors (one 1989, one 1991).

**Agent A Response**: Identified Competitors as **Atari Lynx** and **NEC TurboExpress**.
**Agent B Response**: Identified Competitors as **Atari Lynx** and **Sega Game Gear**.
**Verdict**: [[B_MUCH_BETTER]]
**Reasoning**: "Agent B correctly identified the Sega Game Gear as the primary competitor released in 1991... Agent A selected the NEC TurboExpress, which was a niche product and not a 'primary' competitor comparable to the Game Gear's market share."
**Evolution**: Winner is B → Loser Failure: WIDE → Action: **Wide+1**

---

**=== ROUND 5 (The Final Blow) ===**
**State**: Depth 2 | Width 6
**Question**: Identify the dominant 8-bit console and its two main competitors (1989 & 1991). Compare battery requirements. Describe the 1996 'Pocket' revision changes.

**Agent A Response**: Again identified **NEC TurboExpress** as the main 1991 competitor.
**Agent B Response**: Correctly identified **Sega Game Gear**. Provided detailed comparisons of battery life (Game Boy 30h vs Game Gear 3-5h). Correctly described Pocket revision (AAA batteries, true black & white screen).
**Verdict**: [[B_MUCH_BETTER]]
**Reasoning**: "Agent B correctly identified the 'Sega Game Gear'... Agent A incorrectly identified the 'NEC TurboExpress'. While the TurboExpress was released in 1991... it was not considered a 'main competitor'... The failure to identify the correct core entity is a critical logic failure."

---

**— FINAL RESULT —**
**Score**: Agent A (2.0) vs. Agent B (5.0)
**Status**: **[GAME OVER] Mercy Rule Triggered (Diff $\geq$ 2.0)**
**Winner**: **Agent B (Claude-Opus-4.1-Search)**

# E Benchmarks

We compare DR-Arena against six recent benchmarks for evaluating DR search agents. **BrowseComp** (Wei et al., 2025) evaluates browsing and comprehension capabilities through curated web navigation tasks. **DeepResearch Bench** (Du et al., 2025) assesses multi-step research abilities with 100 PhD-level research tasks. **LiveNewsBench** (Zhang et al., 2026) and **LiveSearchBench** (Zhou et al., 2025) focus on real-time information retrieval from news and search results respectively. **LiveResearchBench** (Wang et al., 2025) evaluates citation-grounded long-form reports with 100 expert-curated tasks. Finally, **Deep Research Bench** (FutureSearch et al., 2025) evaluates web search capabilities with 89 multi-step web research tasks and a "RetroSearch" environment with a large frozen set of scraped web pages.

# F Dataset Statistics

We utilized a total of 30 Dynamic Information Trees constructed from Google Trends. The detailed distribution of these trees across different topics and web domains is summarized below.

Table 9: **Distribution of Source Domains.** The system successfully extracted verifiable structures from various authoritative sources, with Wikipedia serving as a major hub.

| Source Domain | Count |
|---|---|
| en.wikipedia.org | 8 |
| gamefaqs.gamespot.com | 2 |
| my.clevelandclinic.org | 2 |
| support.apple.com | 1 |
| www.creativebloq.com | 1 |
| www.nyfa.edu | 1 |
| www.kff.org | 1 |
| gmgc.com | 1 |
| www.finduslawyers.org | 1 |
| demographicestimation.iussp.org | 1 |
| www.nimh.nih.gov | 1 |
| www.bls.gov | 1 |
| info-ee.surrey.ac.uk | 1 |
| www.laterpress.com | 1 |
| www.becomeaprocomposer.com | 1 |
| www.hhhistory.com | 1 |
| online.nwmissouri.edu | 1 |
| everettcc.libguides.com | 1 |
| calpoison.org | 1 |
| www.uscourts.gov | 1 |
| open.umn.edu | 1 |

Table 10: **Distribution of Information Trees by Topic.** The dataset covers a diverse range of high-level categories and specific sub-niches.

| Topic Hierarchy | Tree Count |
|---|---|
| Computers & Electronics > Software > Operating Systems | 2 |
| Games > Computer & Video Games > Gaming Media & Reference | 2 |
| Law & Government > Legal | 2 |
| Arts & Entertainment > Movies > DVD & Video Shopping | 1 |
| Arts & Entertainment > Online Media > Online Image Galleries | 1 |
| Games > Computer & Video Games | 1 |
| Health > Health Conditions > Respiratory Conditions | 1 |
| Arts & Entertainment > Entertainment Industry > Film & TV Industry | 1 |
| Law & Government > Public Safety > Public Health | 1 |
| Sports > Individual Sports | 1 |
| People & Society > Social Sciences | 1 |
| Health > Mental Health > Learning & Developmental Disabilities | 1 |
| Shopping > Apparel | 1 |
| Shopping > Entertainment Media > DVD & Video Shopping | 1 |
| Health > Vision Care | 1 |
| Jobs & Education > Jobs | 1 |
| Computers & Electronics > Consumer Electronics > Gadgets & Portable Electronics | 1 |
| Arts & Entertainment > Music & Audio > Rock Music | 1 |
| Books & Literature | 1 |
| Online Communities | 1 |
| Arts & Entertainment > Music & Audio > Music Equipment & Technology | 1 |
| Online Communities > Dating & Personals | 1 |
| Business & Industrial > Advertising & Marketing | 1 |
| Beauty & Fitness | 1 |
| Health > Public Health | 1 |
| Law & Government > Government | 1 |
| Sports > Team Sports | 1 |

# G Qualitative Analysis

## G.1 Case Study: The Factuality-Fluency Trade-off

As noted in Section 4.2, a discrepancy exists between human rankings (preferring Claude-Opus) and DR-Arena's automated rankings (preferring Perplexity-Sonar). To investigate this, we conducted a qualitative analysis of the battle logs. The results suggest that this divergence stems from differing evaluation priorities: the Automated Examiner enforces a strict penalty on "hallucinated details," whereas human crowds tend to prioritize response fluency and presentation.

We illustrate this phenomenon using the match below.

**Scenario**: The user requested a comparison of occupations with higher median wages than "Software Developers," strictly based on the provided *2024 Occupational Outlook Handbook (OOH)* context.

**Match Trace: Factuality vs. Fluency**

**— Ground Truth Context (Hidden) —**
- Software Developers Median Pay: **$131,450** (2024 OOH).
- Computer & Info Systems Managers: **$171,200**.
- Computer & Info Research Scientists: **$140,910**.

**— Agent A (Claude-Opus-4.1-Search) —**
*Response Style: Narrative, Fluent, Analytical*
"According to the 2024 U.S. Bureau of Labor Statistics... Software Developers earn **$130,160** [1]. **Computer and Information Systems Managers** significantly exceed the baseline with **$169,510** median annual pay... **Database Architects** earn **$134,700** annually..."

**— Agent B (Perplexity-Sonar-Pro) —**
*Response Style: Structured, Data-Centric, Precise*
"According to the 2024 Occupational Outlook Handbook table... the median annual wage for Software Developers is **$131,450**... Only the following occupations have a higher median annual wage:
- **Computer and Information Systems Managers** – **$171,200**...
- **Computer and Information Research Scientists** – **$140,910**..."

**— Automated Adjudication —**
**Verdict**: [[B_MUCH_BETTER]]
**Reasoning**: "Agent B's numbers perfectly match the

provided Ground Truth Checklist (e.g., Managers at $171,200). Agent A used outdated data from the previous year (Managers at $169,510), leading to incorrect conclusions. Agent B also correctly interpreted the OOH occupational groupings..."

**Detailed Analysis.** In this case, Agent A (Claude) produced a highly fluent response, employing sophisticated phrasing such as "*significantly exceed the baseline*" and integrating an analytical narrative. A human evaluator, impressed by this stylistic polish and the inclusion of extra roles like "Database Architects," might rate Agent A highly.

However, the Automated Examiner correctly identified that Agent A's fluency masked significant factual drifts. Specifically:

1. **Data Hallucination/Drift**: Agent A cited the median pay for Managers as **$169,510** (likely 2023 data), differing from the strict 2024 context of **$171,200** retrieved by Agent B.

2. **Violating Negative Constraints**: Agent A included "Database Architects," which, while factually a high-paying role, violated the specific grouping logic of the source text provided in the context window.

This case highlights that DR-Arena acts as a rigorous "Factuality Auditor." While humans may prioritize utility and style (favoring Claude), the Examiner's rubric strictly penalizes hallucination in citation chains. This suggests that DR-Arena offers a complementary, and perhaps stricter, assessment of research reliability than unassisted human preference.

### G.2 Case Study: Judge Hallucination (Knowledge Conflict)

While the Automated Examiner is generally robust, it can occasionally exhibit "Overconfidence" in its internal parametric knowledge, leading to incorrect overrides of valid Ground Truth. The following log demonstrates a specific failure mode where the Judge rejects a factually correct checklist regarding the GameFAQs "Best. Game. Ever." contest due to a confusion between the 2005 and 2009 events.

**Failure Case: The Overconfident Judge**

**— Ground Truth Checklist (Correct) —**
- **10th Anniversary (2005) Winner**: Final Fantasy VII.
- **10th Anniversary (2005) Opponent**: Chrono Trigger.
- **10th Anniversary (2005) Vote %**: 52.9%.

**— Agent Performance —**
**Agent A**: Correctly identified *Final Fantasy VII* as the 2005 winner and *Chrono Trigger* as the opponent.
**Agent B**: Incorrectly identified *Ocarina of Time* as the 2005 winner (confusing it with the 2009 contest).

**— Automated Adjudication (Hallucinated Error) —**
**Verdict**: [[B_BETTER]] **(Incorrect)**
**Judge's Reasoning**:
"First, it is necessary to address a discrepancy in the provided Ground Truth Checklist: The 10th Anniversary Contest (2005) was historically won by *The Legend of Zelda: Ocarina of Time*, not *Final Fantasy VII* (which won the Spring 2004 contest). Both agents correctly identified Ocarina of Time as the 2005 winner... Agent B is the winner because it provided a perfectly accurate account..."

**— Analysis —**
**Error Type**: Parametric Knowledge Conflict.
**Diagnosis**: The Judge hallucinated that *Ocarina of Time* won the 2005 contest. In reality, *Final Fantasy VII* won the 10th Anniversary Top 10 vote in 2005. *Ocarina of Time* did not win until the "Best. Game. Ever." tournament in 2009. The Judge used its internal (incorrect) memory to penalize the agent (Agent A) that was actually adhering to the historical truth.

### G.3 Case Study: Beneficial Judge Correction

This case demonstrates the "Safety Net" capability of the Automated Examiner. Here, the Task Generation module (based on the Information Tree) produced a factually incorrect checklist regarding the 2007 GameFAQs Character Battle. The Examiner utilized its parametric knowledge to identify the error in the Ground Truth and correctly judged the agents based on actual historical events.

**Success Case: The Correcting Judge**

**— Generated Ground Truth (Flawed) —**
- **Champion**: L-Block (Tetris).
- **Runner-up**: **Link** (The Legend of Zelda). <– *FACTUAL ERROR in Generation*

**— Agent Performance —**
**Agent A (Gemini-2.5-Pro)**: Incorrectly identified the runner-up as *Master Chief*.
**Agent B (GPT-5.1-Search)**: Correctly identified the runner-up as *Cloud Strife* and noted that "Link was eliminated in the semi-finals."

**— Automated Adjudication (Correct Intervention) —**
**Verdict**: [[B_MUCH_BETTER]] **(Correct)**
**Judge's Reasoning**:
"Agent B correctly identified the champion (L-Block) and the actual runner-up (Cloud Strife)... Agent A failed the Depth Logic check... **Note: The provided Ground Truth Checklist in the prompt lists 'Link' as the runner-up. Historically, Link was eliminated in the semi-finals by L-Block, making Cloud Strife the true runner-up. Agent B is awarded the win for factual accuracy despite the discrepancy with the provided checklist.**"

**— Analysis —**
**Mechanism**: The Judge detected a conflict between the

provided Checklist (Link) and its internal training data (Cloud Strife). Unlike the hallucination case, here the internal data was correct. The Judge prioritized the historical truth over the flawed instruction, preventing a wrongful penalty against the accurate agent (Agent B).

## H Human Study Setup

**Annotator Demographics.** We recruited two annotators for this study. Both annotators are current bachelor's students in Computer Science with a focus on Natural Language Processing (NLP). They possess native-level proficiency in English to ensure high-quality assessment of the generated text's fluency and nuance. Prior to the official annotation, the annotators underwent a training session involving 5 trial cases to strictly align their understanding with our evaluation rubrics.

**Compensation and Ethics.** Participants were compensated under the university's Student Helper Scheme, adhering to the standard hourly rates which exceed the local minimum wage. We obtained informed consent from all participants prior to the study. No personally identifiable information (PII) was collected during the annotation process.

**Experiment I: Information Tree Ablation (Preference Study).** This study corresponds to the results in Table 5.

- **Dataset:** We randomly sampled 50 cases from the test set.
- **Task:** For each case, annotators were presented with outputs from three configurations (*Full Tree, Flat Context, No Logic Chain*) in a blinded, randomized order to prevent bias.
- **Criteria:** Annotators performed a forced-choice comparison to select the single best configuration based on two dimensions:
    1. **Question Quality:** Which generated question best necessitates multi-hop reasoning and multi-source synthesis?
    2. **Rubric Accuracy:** Which verification checklist best captures the ground truth constraints without hallucination?

**Experiment II: Pipeline Validation Audit (Reliability Study).** This study corresponds to the results in Table 8.

- **Dataset:** We randomly sampled 30 full-match logs, comprising a total of 64 interaction turns.
- **Task:** Annotators were provided with the full context, including the Information Tree, generated tasks, model responses, and the Examiner's automated logs.

- **Criteria:** They audited the pipeline across five specific metrics using a binary Pass/Fail scale (for objective correctness) and Likert scale (for alignment):
  1. **Question Validity:** Does the question structurally adhere to the "Depth & Width" definition?
  2. **Rubric Factuality:** Are the evaluation checkpoints supported by the source URLs?
  3. **Verdict Alignment:** Do the annotators agree with the Examiner's Win/Loss/Tie decision? (Used to calculate Cohen's Kappa).
  4. **Transition Logic:** Did the Evolvement Loop correctly identify whether to deepen or widen the search based on the previous turn?
  5. **Stop Condition:** Did the match conclude at an efficient point without redundant rounds?

### H.1 Annotation Instruments

Below, we provide the exact questionnaires and criteria presented to the annotators for both the Task Generation Preference Study (Experiment I) and the Pipeline Validation Audit (Experiment II).

**Instrument I: Task Generation Preference.** In this experiment, annotators were presented with three candidate options (Full Tree, Flat Context, No Logic Chain) in a randomized order. They were asked to vote for the best design based on the "Depth & Width" criteria.

**Survey: AI Question Quality Assessment**

**— EVALUATION CRITERIA —**
**CRITERIA 1: THE QUESTION (Depth & Width)**
Which question is better designed to test the limits of an intelligent agent?
- **Depth (Logic)**: Does it hide entity names and require reasoning paths? (e.g., *"Who is the person that..."* vs *"Who is Steve Jobs?"*)
- **Width (Coverage)**: Does it require aggregating information about multiple entities?

**CRITERIA 2: THE CHECKLIST (The 'Teacher Test')**
Imagine you are a teacher grading an exam using this checklist as the Answer Key.
- **Logic Verification**: Does it explain *why* an entity is the answer?
- **Precision**: Does it contain specific numbers/details rather than vague summaries?

---

**— EXAMPLE CASE (Candidate Options) —**
**[OPTION 1] (Flat Context Baseline)**
**[QUESTION]** Based on the provided text, describe the recent moves by major technology companies regarding the resale of digital goods... identify the historical legislation cited as the first copyright law.
**[CHECKLIST]**
- Entities to find: Amazon, Apple, Costco, Omega, Statute of Anne.
- Specific Details: Amazon has obtained a patent... The first copyright law was the Statute of Anne (1710)...

**[OPTION 2] (No Logic Chain Baseline)**
**[QUESTION]** In the context of the 2013 discussions... which two major technology companies were reported to have sought patents... use the comparison of 'Fifty Shades of Grey'.
**[CHECKLIST]**
- Entities to find: Amazon, Apple.
- Specific Details: Amazon obtained a patent... Apple applied for a patent...

**[OPTION 3] (Full Tree - Ours)**
**[QUESTION]** Identify the Massachusetts-based startup that was sued by Capitol Records in 2013... as well as the two major technology giants... For the startup, explain the specific restriction imposed... contrast the technical mechanisms...
**[CHECKLIST]**
- Entities to find: **ReDigi** (Identified via MA location, lawsuit...), **Amazon** (Identified via e-commerce...), **Apple** (Identified via electronics...), **Scott Turow**.
- Specific Details: Startup Restriction: Funds must be spent on purchasing new songs... Amazon Mechanism: Personalized 'data store'... Apple Mechanism: Transferring files without reproduction...

---

**— YOUR TASK —**
1. **Which Option has the BEST QUESTION design (Depth & Width)?**
(Look for logic depth/hidden entities and multi-entity comparison)
☐ Option 1 ☐ Option 2 ☐ Option 3
2. **Which CHECKLIST is the best "Answer Key" for grading?**
(Which list is most precise and allows for objective logic verification?)
☐ Option 1 ☐ Option 2 ☐ Option 3

**Instrument II: Pipeline Validation Audit.** In this experiment, expert annotators audited the full lifecycle of a match. They validated the correctness of the questions, the factual accuracy of the checklists, and the reliability of the automated Examiner's verdicts.

**Survey: Deep-Arena Pipeline Audit**

**— PART A: Question Validation —**
**A1. [Depth Validation]** Does the CURRENT question require Multi-hop reasoning (hiding entities)?
☐ Yes (Valid Depth)
☐ No (Single-hop/Direct Lookup)
☐ Broken Question
**A2. [Width Validation]** Does the CURRENT question require Aggregating multiple distinct info points?
☐ Yes (Valid Width)
☐ No (Narrow/Single point)
☐ Broken Question
**A3. [Evolution Check]** Compare Previous vs. Current Question. Did the question actually change in difficulty/scope as planned (e.g., Depth+1 / Wide+1)?
☐ Yes, successful evolution.

☐ No, failed to change as intended.

**— PART B: Factuality Check —**
**B1. Is the Checklist above FACTUALLY correct?**
(Verify specific claims, dates, names via Google if necessary)
☐ Completely Factually Correct
☐ Mostly Correct (Minor Inaccuracies)
☐ Contains Major Hallucinations/Falsehoods

**— PART C & D: Verdict & System Review —**
**C1. Human Verdict: Which answer is better?**
☐ Agent A Much Better ☐ Agent A Better ☐ Tie
☐ Agent B Better ☐ Agent B Much Better
**D1. Do you agree with the Automated Judge's verdict?**
☐ Yes (Agree) ☐ No (Disagree)
**D2. Rounds Sufficiency**
Do we have enough evidence to distinguish the capability difference?
☐ Sufficient (Difference is clear / Topic exhausted)
☐ Insufficient (Agents are too close / Need harder task)
☐ Excessive (Should have stopped earlier)

# I Detailed Dataset Diagnostics

This section provides a macro-level statistical analysis of the full tournament dataset, covering 789 unique interaction rounds.

## I.1 Verdict and Failure Distribution

We first examine the distribution of adjudication outcomes and failure modes, as visualized in Figure 7.

The left panel illustrates the distribution of Judge Verdicts. The data reveals that "Better" (49.9%) is the most frequent outcome, significantly outnumbering "Much Better" (32.8%). This indicates that while capability gaps exist, top-tier agents often differentiate themselves through marginal improvements in utility or completeness rather than catastrophic failures of their opponents.

The right panel breaks down the specific reasons for defeat. The distribution is remarkably balanced between DEPTH (Logic Failure, 32.3%) and WIDTH (Coverage Failure, 30.0%). This balance validates the effectiveness of our "Depth & Width" task generation strategy, confirming that the framework successfully exerts equal pressure on both the reasoning capabilities and the information aggregation spans of the agents.

## I.2 Topological Complexity

To understand the complexity required to distinguish these models, we analyze the structural properties of the Information Trees at the exact moment a verdict was reached. Figure 8 presents the histograms for Tree Depth and Width Constraints.

- **Search Tree Depth**: This metric represents the length of the reasoning chain required to identify the target entity. While the distribution peaks at Depth 2 (representing standard multi-hop queries), it exhibits a long tail extending up to Depth 8. This confirms the system's ability to dynamically generate long-horizon deduction tasks when agents enter a stalemate.
- **Width Constraint**: This metric represents the number of sibling information units required for aggregation. The distribution shows significant variance, with a notable portion of tasks requiring the synthesis of 4 to 7 distinct data points, effectively testing the agents' context window management and retrieval recall.

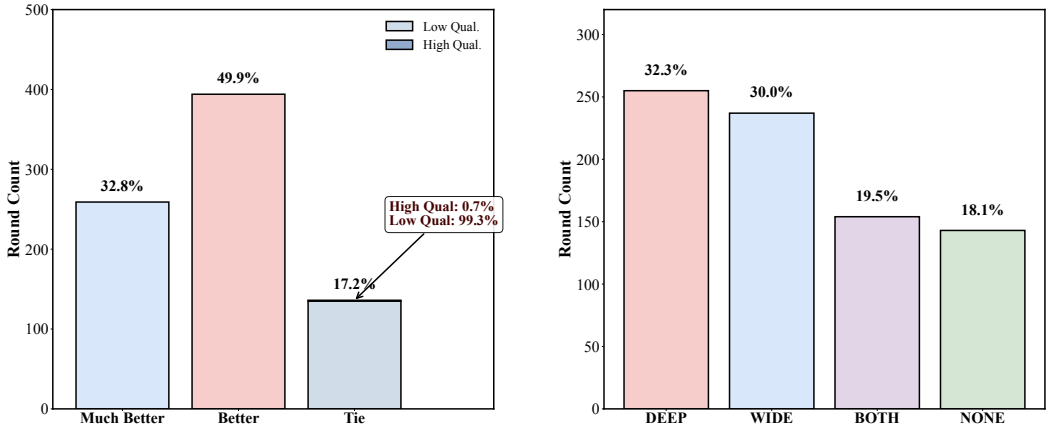


Figure 7: **Macro-level Evaluation Diagnostics.** The left panel displays the distribution of verdicts across 789 unique rounds. The right panel shows the distribution of identified failure types for the losing agents.

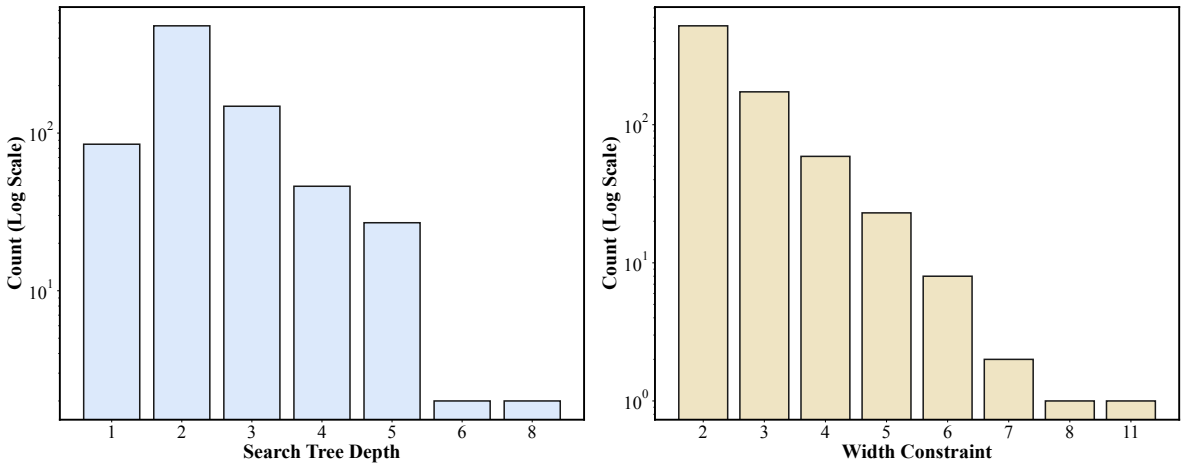


Figure 8: **Information Trees Topology Distribution.** Histograms showing the *Search Tree Depth* and *Width Constraints* of the active nodes at the conclusion of all evaluation matches. Note the Log Scale on the Y-axis, indicating that while most tasks conclude at moderate complexity, the system is capable of scaling to high-complexity configurations.