Agentic Search in the Wild: Intents and Trajectory Dynamics from 14M+ Real Search Requests

Jingjie Ning*
jening@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, US

João Coelho*
jmcoelho@andrew.cmu.edu
INESC-ID, Carnegie Mellon University
Pittsburgh, PA, US

Yibo Kong*
yibok@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, US

Yunfan Long*
justinlo@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, US

Bruno Martins
bruno.g.martins@tecnico.ulisboa.pt
INESC-ID, Instituto Superior Técnico, University of Lisbon
Lisbon, Portugal

João Magalhães
jm.magalhaes@fct.unl.pt
NOVA LINCS
NOVA University Lisbon
Caparica, Portugal

Jamie Callan
callan@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, US

Chenyan Xiong
cx@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, US

Abstract

LLM-powered search agents are increasingly being used for multi-step information seeking tasks, yet the IR community lacks empirical understanding of how agentic search sessions unfold and how retrieved evidence is reflected in later queries. This paper presents a large-scale log analysis of agentic search based on 14.44M search requests (3.97M sessions) collected from DeepResearchGym, i.e., an open-source search API accessed by external agentic clients. We sessionize the logs, assign session-level intents and step-wise query-reformulation labels using LLM-based annotation, and propose Context-driven Term Adoption Rate (CTAR) to quantify whether newly introduced query terms are lexically traceable to previously retrieved evidence. Our analyses reveal distinctive behavioral patterns. First, over 90% of multi-turn sessions contain at most ten steps, and 89% of inter-step intervals fall under one minute. Second, behavior varies by intent. Fact-seeking sessions exhibit high repetition that increases over time, while sessions requiring reasoning sustain broader exploration. Third, query reformulations are often traceable to retrieved evidence across steps. On average, 54% of newly introduced query terms appear in the accumulated evidence context, with additional traceability to earlier steps beyond the most recent retrieval. These findings provide candidate signals for repetition-aware stopping, intent-adaptive retrieval budgeting, and explicit cross-step context tracking. We released the anonymized logs, making them available at a public HuggingFace repository.

CCS Concepts

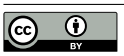
• **Information systems** → **Query log analysis**.

Keywords

Agentic Search, Query Log Analysis, Deep Research, Search Intent

ACM Reference Format:
Jingjie Ning, João Coelho, Yibo Kong, Yunfan Long, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong. 2026. Agentic Search in the Wild: Intents and Trajectory Dynamics from 14M+ Real Search Requests. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809627

*These authors contributed equally to this work.



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809627

1 Introduction

Information retrieval is shifting from human-initiated search into agentic search [2, 31, 41, 52], where LLM-powered agents plan and execute multi-step information seeking with retrieval tools. Instead of issuing a single query and consuming a ranked list, an agent may iteratively reformulate queries, retrieve evidence, and issue later queries in response to the returned context. While agent capabilities are increasingly demonstrated on controlled benchmarks [20, 30, 50], benchmark scores alone do not reveal how agents' queries evolve across steps, or how context is reflected in later queries.

These questions matter for practical system design. Agents may spend retrieval budget on repetitive or overly narrow reformulations, fail to explore alternative facets, or carry forward little useful context across steps. Understanding session structure can inform query-policy control, and measuring evidence traceability can guide budget allocation and evaluation design. As agents consume results programmatically, leaving no direct trace of what they found useful, logs lack implicit feedback signals, such as the clicks that anchor traditional behavioral inference. This creates a measurement gap. We can observe sequences of submitted queries and returned evidence, but it remains unclear how sessions unfold, how behavior differs by intent, what reformulation moves dominate, and whether later queries are lexically traceable to evidence returned earlier.

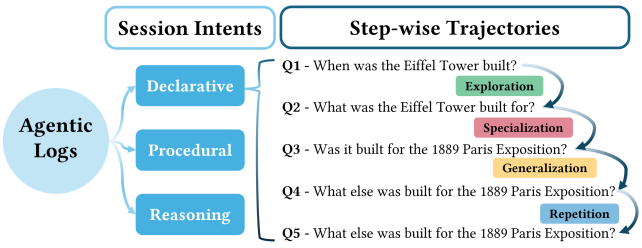


Figure 1: Intent–trajectory structure of agentic search logs.

To address this gap, we analyze agentic search at two complementary levels, namely what the agent is trying to accomplish in a session, which we capture with session-level intent, and how the agent pursues that goal through step-wise search actions, which we capture with trajectory-level query reformulation. Figure 1 illustrates this structure on a short example session. To operationalize these levels, we develop a measurement framework with three components. First, we use LLM-based annotation pipelines to assign interpretable intent and trajectory labels to sessions and step-pairs, following standard taxonomies. Second, we replay logged queries offline to reconstruct the evidence returned at each step. Third, we introduce Context-driven Term Adoption Rate (CTAR), a metric that quantifies whether newly introduced query terms can be lexically traced to retrieved evidence, including traceability to steps beyond the most recent retrieval.

We apply this framework to logs collected from DeepResearchGym (DRGym) [7], a research-oriented reproducible search API accessed by external agentic clients. With permission from the DRGym organizers, we study 14.44M logged search requests spanning six months, which we sessionize into 3.97M sessions. The resulting data provides an at-scale view of autonomous agents operating in the wild under a shared retrieval backend, while still preserving the distinction between observable API-level traces and unobserved client-side prompts, memory, and control policies.

Our results characterize retrieval budgeting, reformulation behavior, and evidence traceability in multi-step agentic search. We find that over 90% of multi-turn sessions contain at most ten steps, and 89% of inter-step intervals fall under one minute. Retrieval depth, measured by the number of documents requested per query, is largely static, suggesting that many clients treat it as a fixed parameter rather than adapting it within sessions. Behavior also varies by intent. Fact-seeking sessions exhibit the highest repetition, which increases over time, showing that near-duplicate loops can emerge in later steps, while sessions requiring reasoning sustain broader exploration throughout. We also find that many newly introduced query terms are lexically traceable to previously retrieved evidence, with measurable overlap from earlier steps beyond the most recent retrieval.

Our contributions can be summarized as follows:

- We provide a large-scale behavioral characterization of agentic search from a reproducible search infrastructure (14.44M search requests, 3.97M sessions), offering an at-scale view of autonomous agents operating in the wild.
- We introduce CTAR, as a metric for quantifying evidence-conditioned query evolution, and use it to measure cross-step lexical traceability beyond the most recent retrieval.
- We identify candidate design signals from the logs, including repetition-aware stopping, intent-adaptive retrieval budgeting, and cross-step context tracking.

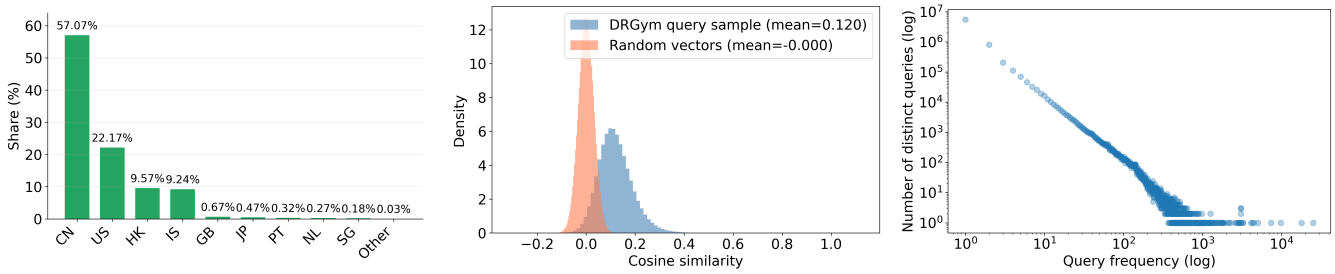
We have released the anonymized logs to support future research and reproducibility, making them available as a public HuggingFace dataset repository.

2 Related Work

Human Search Behavior and Log Analysis: Large-scale query logs have long been used to study search behavior in the wild [8, 17, 42], offering scalable, behavior-grounded signals for characterizing session dynamics and query reformulation beyond what offline benchmarks capture. A core theme is within-session learning. Eickhoff et al. [10] trace how newly introduced terms relate to evidence observed before reformulation (e.g., SERP snippets and visited pages), alongside complementary work on interpreting implicit feedback such as clicks and dwell time [1, 12, 21]. Exploratory search and navigation studies further document differences in branching and interaction patterns across users and information needs [29, 44, 49], while sessionization analyses examine how sessions begin and end in practice [13, 22, 42]. We adopt this evidence-traceability perspective for autonomous agents and operationalize it using retrieved evidence text, enabling systematic comparisons of agent behaviors across intents and guiding design choices such as retrieval budgeting and cross-step context management. While a small line of work compares humans and agents directly [47, 48, 57], these comparisons are often task-specific or simulation-based, motivating complementary large-scale log analyses of how autonomous agents search across sessions in the wild.

LLM Interaction Platforms and Usage Logs: Recent efforts analyze large-scale interaction data from LLM systems and evaluation platforms. Chatbot Arena (LMSYS LLM Arena) aggregates pairwise preference votes [6], and LMSYS-Chat-1M releases one million multi-model conversations collected in the wild [56]. OpenAI reports how people use ChatGPT at scale [34], and Anthropic presents privacy-preserving analyses of millions of Claude conversations to characterize economic task usage [15]. SciArena extends the Arena protocol to scientific literature-grounded tasks and provides a corresponding benchmark [54]. These works capture usage and preference signals, but typically do not expose tool-level retrieval traces (queries, evidence, step-wise search decisions) needed to study agentic search behavior and within-session evidence reuse.

Agentic Search Modeling, Benchmarks, and Infrastructures: Recent systems enabling LLMs to plan multi-step interactions with retrieval tools have shifted IR toward agentic workflows [2, 31, 41, 52]. Benchmarks for tool-using agents include WebShop [51], WebArena [58], AgentBench [28], and large-scale tool-use evaluation such as ToolLLM/ToolBench [38]. DeepResearchGym (DRGym) provides an open-source sandbox with a reproducible search API and evaluation protocol for deep research systems [7]. Early analyses have begun to formalize agent behaviors. Jin et al. [20] link beneficial reasoning patterns to gains on GAIA [30] and WebWalker [50];



(a) Geographic distribution of requests. **(b) Pairwise query similarity (100k sample).** **(c) Query frequency distribution (log–log).**

Figure 2: Representativeness and diversity of the DRGym logs.

complementary efforts propose taxonomies and risk frameworks, such as ST-WebAgentBench [25] and the Agentic AI Security Scoping Matrix [5]. However, benchmark scores alone provide limited visibility into how agents search in practice, and prior human–agent comparisons suggest differences in query breadth and context use that are difficult to diagnose without session-level traces [47, 48, 57]. Most prior work focuses on benchmarking and system design, whereas we measure behavior at scale from real logs, and quantify evidence traceability via CTAR.

3 Data and Log Processing

In this section, we start by providing an overview of DRGym to better contextualize our analysis. Then, we describe the query log we have been given access to, following with a presentation of the preprocessing and session segmentation (sessionization) pipeline used to convert raw requests into sessions.

3.1 DRGym Log Overview

DRGym serves requests from external agentic clients, capturing diverse usage patterns and interaction styles. The API is model-agnostic, i.e., it operates as a retrieval backend rather than a single deployed agent. The logs therefore do not include the client-side model, prompting strategy, memory policy, or the way retrieved evidence is shown to the agent. Requests nevertheless share the same retrieval infrastructure, which lets us study API-level traces, including submitted queries, request parameters, timestamps, and evidence returned by the backend.

The backend performs dense retrieval [18, 23] over two large-scale English corpora, i.e. *ClueWeb22-A-EN* [35] and *FineWeb* [36]. The DRGym paper describes a retrieval API with a `/search` endpoint for ranked retrieval [7]. Operating over static web snapshots, rather than a changing live index, enables re-issuing queries under fixed corpora for consistent retrieval behavior across experiments.

Consistent with this design, each log entry records the timestamped query and request parameters (Table 1), including retrieval depth and an ANN search-budget parameter [18]. For privacy, IP addresses are anonymized and used only for coarse client-level aggregation (grouping, sessionization, and country-level reporting), and never for user identification or fine-grained geolocation.

Scale and Coverage: The logs span 2025-06 to 2025-12 and contain 14.44 million requests. After preprocessing and sessionization (Section 3.2), we obtain 3.97M sessions. Requests originate from 558 anonymized client IPs across 25 countries. Figure 2a shows the geographic coverage, with the largest shares from China and the United States, followed by Hong Kong and Iceland. Overall traffic is substantial, peaking at 2.49 million requests in a single week.

Table 1: Fields recorded in the `search_logs` table.

Field	Description
`id`	Auto-incremented unique identifier for each request.
`ip_address`	Client IP address (anonymized for analysis).
`query_text`	Query string submitted to the API.
`num_of_docs`	Number of documents requested for retrieval.
`complexity`	ANN retrieval complexity controlling search budget.
`dataset`	Corpus used for retrieval (ClueWeb22 or FineWeb).
`timestamp`	Timestamp of the logged request.

Semantic Diversity: Figure 2b plots the pairwise cosine-similarity distribution for 100k randomly sampled queries using Qwen3-Embedding-0.6B [53]. The distribution (mean=0.12) lies close to the random-vector baseline (mean≈0 for uniformly distributed vectors), indicating that queries are semantically diverse rather than clustered around repeated themes. The slight rightward shift likely reflects shared information-seeking phrasing, not semantic redundancy. For reference, Qwen3 uses cosine similarity > 0.7 to mark semantically related pairs during training [53].

Query-Level Repetition: Figure 2c shows a long-tailed query frequency distribution. Most distinct queries are quite rare, while only a small set of queries repeats often. In particular, 53.89% of distinct queries occur at most three times, including 38.38% singleton queries, and the top-10 and top-100 most frequent queries account for only 0.59% and 1.51% of all requests, respectively.

Taken together, the broad geographic coverage, low average semantic similarity, and long-tailed frequency spectrum suggest that the DRGym query stream contains a diverse mixture of information needs, rather than a narrow set of repeatedly executed prompts.

As an additional sanity check, we estimate whether the stream is concentrated around several widely used agentic and deep-research benchmarks. To that end, we measure semantic overlap between log queries and four widely used benchmarks: GAIA [30], FRAMES [24], HLE [37], and WebWalkerQA [50]. These benchmarks provide a check against concentration on common deep-research, reasoning, and web navigation tasks. We consider a sample of 1 million queries

Table 2: Semantic overlap between selected agentic benchmarks and a 1M log query sample (cosine similarity ≥ 0.7).

Benchmark	Bench. Queries	Hits in Sample	% of Sample
GAIA	103	27	0.00%
FRAMES	824	879	0.09%
HLE	2,158	616	0.06%
WebWalkerQA	680	2,219	0.22%
Total	3,765	3,741	0.37%

from the logs, and encode all benchmark queries using Qwen3-Embedding-0.6B. Then, we count log queries exceeding a cosine similarity threshold of 0.7 to any benchmark query. As shown in Table 2, benchmark-similar queries constitute less than 0.4% of the sample across all four benchmarks combined. Overlap is lowest for GAIA, and highest for WebWalker, whose web traversal queries more closely resemble natural search formulations. Overall, these results suggest that the logs reflect diverse open-ended usage rather than concentrated benchmark execution.

To support reproducibility and enable further research, we have released the cleaned and anonymized logs associated with this study at a public Hugging Face dataset repository. We have removed direct identifiers (e.g., IP addresses) and applied standard PII scrubbing on free-text fields, releasing only the fields needed to reproduce our analyses with anonymized session IDs. We have documented the anonymization procedure and residual risks in the dataset card, following prior large-scale LLM interaction log releases and privacy-preserving analyses [15, 56]. After the initial publication, the dataset may be updated as additional logs are collected and validated.

3.2 Log Preprocessing and Sessionization

We first remove malformed entries (e.g., empty queries), internal testing traffic, and outlier repetition bursts, before segmenting the remaining stream into sessions.

Although standard sessionization often relies on fixed time-gap heuristics, agentic requests can arrive in fast parallel patterns [32], making a pure temporal cutoff unreliable. We therefore sessionize with a semantic-continuity criterion combined with an explicit temporal constraint. Concretely, for each IP we maintain active sessions and assign an incoming query to the most semantically continuous active session when the continuity score exceeds a threshold; otherwise we start a new session. We additionally impose a 10-minute hard cutoff between consecutive queries within a session, reflecting faster interaction loops than the conventional 30-minute rule for human logs [22, 42]. Among classifier-predicted continuous pairs, only 0.92% have gaps exceeding 10 minutes.

The aforementioned pipeline yields 3.97M sessions. Manual spot-checks confirm that the resulting sessions are generally coherent. Full procedural details (i.e., continuity model, thresholds, and validation) are provided in Appendix A.

4 Methodology

To address the questions motivating this work, we require measurements at two levels, namely session-level intent, which captures the type of information need driving a session, and trajectory-level reformulation, which captures how queries change from step to step. We also require a way to assess whether later queries are lexically traceable to evidence returned earlier in the session. For intent and trajectory labeling, we use standard LLM-as-a-judge pipelines [26, 55]. For evidence traceability, we introduce a new metric, namely the Context-driven Term Adoption Rate (CTAR).

We segment the log into sessions $\mathcal{S}$, where each session $s = (q_1, \ldots, q_{|s|})$ is an ordered sequence of timestamped queries. Retrieval depth is denoted by K, corresponding to the logged parameter `num_of_docs`. We analyze behavior at three granularities: global (corpus-wide), session-level (intent-conditioned), and trajectory-level (adjacent query pairs within a session $q_k \rightarrow q_{k+1}$).

4.1 LLM-based Intent and Trajectory Labeling

Session-Level Intent: Different information needs may induce different search strategies. For instance, a user seeking a factual answer may behave differently from one debugging a procedure or reasoning through a complex question. To test whether agentic search exhibits such intent-conditioned structure, we label each session with an intent category. We adopt a three-way taxonomy from web search goal modeling [4, 10, 40] corresponding to the following classes: Declarative (fact retrieval), Procedural (method execution), and Reasoning (complex synthesis). Since q_1 is often already a reformulation, we assign intent from the whole session.

Trajectory-Level Reformulation: Intent alone does not reveal how agents iterate within a session. An agent might narrow its query, broaden it, pivot to a related facet, or retry with a near-identical phrasing. These reformulation patterns have implications for retrieval efficiency: excessive repetition wastes budget, while a lack of exploration may leave relevant facets unexamined. To capture these dynamics, we label each adjacent query pair ($q_k \rightarrow q_{k+1}$) with a trajectory type grounded in prior reformulation taxonomies [3, 16]: Specialization (narrowing by adding constraints), Generalization (broadening by relaxing constraints), Exploration (within-topic facet pivots) and Repetition (identical or near-duplicate reformulations). Representative examples are provided in Appendix C.

Implementation. We implement labeling with `gpt-5-nano` [33]. We annotate multi-turn sessions with $|s| \in [2, 10]$ for intent (one label per session) and all adjacent pairs for trajectories (one label per pair). We focus on this range because our sessionization analysis (Section 5) reveals that it covers 90.32% of all multi-turn traffic, representing the core behavior of current agents. For sessions with mixed or ambiguous signals, the judge assigns the dominant intent, and we interpret intent-conditioned results as aggregate trends rather than definitive labels for every individual session. To assess labeling robustness, we compare labels from two models (`gpt-5-nano` and `gemini-3-flash-preview` [14]) on a 2000-pair random subset, achieving 95.15% agreement. The remaining disagreements are spread across categories rather than concentrated in any single label. Prompts are provided in Appendix D. Unless otherwise noted, analyses from Section 6 onward use a random subset of labeled multi-turn sessions under the annotation budget, excluding single-query sessions and outlier long-tail sessions as described in Section 5. We also compute auxiliary metrics used throughout the paper, each defined at first use with the summary shown in Appendix B.

Table 3: Session-level descriptive statistics by intent type. Formulas for less standard metrics are given in Appendix B.

Statistic	Metric (*Sample N / Ratio*)	Declarative (fact-seeking) (99.8k / 88.64%)	Procedural (how-to) (4.5k / 3.96%)	Reasoning (analytical) (8.3k / 7.41%)
Mean	Session Length	4.03	3.81	4.03
	Retrieval Depth (K)	7.70	37.34	24.99
	Query Length (whitespace terms)	7.59	10.58	12.69
	Initial-Final Gap	0.21	0.22	0.28
Median	Total Duration (s)	40.00	26.00	31.00
	Step Latency (s)	17.00	13.00	14.00

4.2 Context-driven Term Adoption Rate (CTAR)

Agentic search iterates between retrieval and query formulation. Evidence returned at step k may be reflected in later query reformulations. Yet agentic logs provide no direct signal of what the agent actually attended to in retrieved documents, making evidence use hard to observe. We therefore ask a more tractable traceability question: when the agent introduces new query terms at step k+1, do those terms appear in evidence returned before that step? This aligns with evidence-traceability perspectives in human log analysis [10] and searching-as-learning research [9, 39, 46], but has not been systematically studied for autonomous agents.

We formulate this idea through Context-driven Term Adoption Rate (CTAR), i.e. the fraction of newly introduced query terms that can be lexically traced to retrieved evidence. We use exact-match tracing rather than semantic similarity because it is interpretable without threshold tuning, robust across domains and query styles, and conservative, i.e., semantic variants would typically yield higher rates by crediting paraphrases and near matches.

Let $Terms(x)$ denote the set of unique, lowercased, and non-stopword tokens that can be extracted from text x. For a trajectory $(q_k \rightarrow q_{k+1})$, the set of newly introduced terms is:

$$NewTerms(q_{k+1}, q_k) = Terms(q_{k+1}) \setminus Terms(q_k). \quad (1)$$

Let E_k denote the textual evidence returned by the DRGym backend at step k. Since raw logs do not store retrieved documents, we reconstruct E_k by querying the DRGym API using the original logged parameters. Because client-side prompts and memory are not logged, E_k should be interpreted as retrievable evidence exposed through the API, rather than a direct observation of what the agent attended to or retained. We consider two context definitions:

$$C_k^{\text{last}} = Terms(E_k), \quad (2)$$

$$C_k^{\text{agg}} = \bigcup_{i=1}^{k} Terms(E_i). \quad (3)$$

CTAR is the fraction of new terms appearing in the chosen context:

$$\text{CTAR}_k^{(\cdot)} = \frac{\left|NewTerms(q_{k+1}, q_k) \cap C_k^{(\cdot)}\right|}{|NewTerms(q_{k+1}, q_k)|}, \quad (\cdot) \in \{\text{last}, \text{agg}\}. \quad (4)$$

In the previous equation, if $NewTerms(q_{k+1}, q_k) = \emptyset$ or $C_k^{(\cdot)} = \emptyset$, we set $\text{CTAR}_k^{(\cdot)} = 0$. In summary, CTAR quantifies the degree to which query evolution is lexically traceable to returned evidence. CTAR^{last} captures traceability to the immediately preceding step, while CTAR^{agg} captures traceability to evidence returned at any prior step. Comparing the two lets us measure how much additional lexical overlap is introduced by aggregated context, without assuming that the agent causally used or retained that evidence.

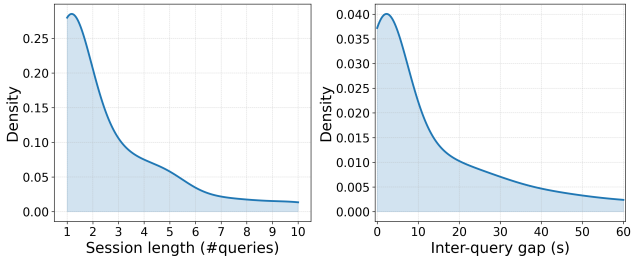


Figure 3: Left: distribution of session length (number of queries per session). Right: distribution of step-wise time intervals between consecutive requests.

5 Aggregate Session Statistics

Session Length and Structural Composition. The corpus comprises 3.97M unique search sessions with a skewed length distribution (Figure 3, left). Nearly half (47.77%) are single-query sessions. Because our trajectory metrics require at least two queries, the rest of the paper focuses on multi-turn sessions. We do not treat single-query sessions as necessarily successful or simple. Instead, our findings characterize sessions that continue beyond one request, which may over-represent complex, uncertain, or iterative information needs. Among multi-turn sessions, 90% have length ≤ 10, indicating that most multi-turn sessions in the log remain short, though some extend considerably further.

For reference, human web search logs are slightly shorter on average (1.7 queries per session) and include a substantially larger fraction of single-query sessions (77.6%) [10, 42]. This comparison suggests that agentic systems often engage in more extended information-seeking episodes, but it should not be interpreted as a direct comparison of task success or difficulty. We focus subsequent analyses on multi-turn sessions.

Temporal Dynamics and Interaction Speed. Within-session intervals are short for most steps, where 56.12% fall within 0–10 seconds, and 89.21% are under one minute (Figure 3, right). We use step latency to denote the wall-clock interval between consecutive API requests. This interval may include client-side LLM inference, batching, network delay, and scheduling overhead, so we treat it as a descriptive pacing signal rather than a direct measure of agent

Table 4: Descriptive statistics by trajectory type. Formulas for less standard metrics are given in Appendix B.

Statistic	Metric (*Sample N / Ratio*)	Specialization (73.8k / 21.76%) *Narrowing*	Generalization (30.6k / 9.02%) *Broadening*	Exploration (125.8k / 37.07%) *Facet pivots*	Repetition (109.1k / 32.15%) *Near-duplicates*
Mean	Dense Similarity	78.18%	80.07%	55.08%	96.70%
	Jaccard Similarity	40.13%	44.66%	25.10%	81.63%
	Result Overlap	22.58%	23.47%	7.35%	78.23%
	Delta Query Length	+1.47	-2.42	+0.04	-0.08
Median	Step Latency (s)	7.00	8.00	14.00	6.00

deliberation time. Intervals are heavy-tailed, reflecting occasional long-latency steps due to system or pipeline delays.

For reference, prior human log studies report median dwell times of several minutes for knowledge-acquisition intents [10]. While dwell time and session duration are not directly comparable to our step latency, the contrast highlights the faster iterative pacing typical of agentic search.

Retrieval Depth and Parameter Stability: Retrieval depth is concentrated at fixed values $K \in \{1, 5, 10\}$, with only 8.36% of sessions using other values. Furthermore, only 1.35% of sessions vary K across steps. Since DRGym supports $1 \leq K \leq 100$, this suggests that many agents treat retrieval count as effectively hard-coded rather than adapted within a session.

6 Intent-Conditioned Session Behavior

Using the LLM-as-a-judge pipeline described in Section 4, we label each multi-turn session as either being Declarative (fact-seeking), Procedural (how-to or step-by-step tasks), or Reasoning (comparative, analytical or multi-hop questions). Declarative dominates (88.64%), followed by Reasoning (7.41%) and Procedural (3.96%). Table 3 summarizes the session-level behavior, reporting medians for time-based measures due to heavy tails, and mean values for count and semantic measures.

Beyond length and timing, we characterize sessions with two additional measures. Retrieval Depth summarizes per-step K at the session level, and Initial-Final Gap measures semantic drift from first to last query using $1 - \cos(q_1, q_{|s|})$ with Qwen3-Embedding-0.6B [53]. Formal definitions are given in Appendix B.

Declarative sessions use the shallowest retrieval yet incur the highest interaction costs. This pattern is consistent with agents using more iterations when per-step retrieval is shallow, although the logs do not determine whether the extra steps improve verification. The pattern contrasts with human fact-finding, where users issue fewer and shorter queries [8, 10, 42]. Agents instead phrase queries as full constraint-bearing questions, consistent with iterative verification behavior [20].

Procedural sessions show the opposite pattern. Deeper retrieval accompanies a more semantically stable progression, consistent with broader evidence coverage co-occurring with fewer refinement steps. Queries within these sessions are longer than Declarative ones, consistent with prior studies of procedural search that have reported similar characteristics [10].

Table 5: Distribution of trajectory types across all step-wise transitions within sessions of each intent.

Trajectory	Session Intent		
	Declarative	Procedural	Reasoning
Specialization	21.12%	27.99%	26.35%
Generalization	9.07%	10.12%	7.89%
Exploration	36.12%	39.04%	47.57%
Repetition	33.69%	22.85%	18.19%

Finally, we observe that Reasoning sessions match Declarative in turn count but differ in how queries evolve. They show the largest semantic drift and longest queries, while retrieval depth is moderate. The distinguishing signal for Reasoning lies in within-session query reformulation rather than session duration or retrieval depth.

7 Trajectory Moves and Topologies

In this section, we analyze how agents revise queries step by step. This is a defining characteristic of agentic search that exposes intermediate decision-making and supports more fine-grained analysis than single-shot querying.

7.1 Trajectory Types and Properties

We follow the labeling procedure in Section 4.1, classifying each adjacent pair $(q_k \rightarrow q_{k+1})$ as Specialization (narrowing by adding constraints), Generalization (broadening by relaxing constraints), Exploration (within-topic facet pivots), or Repetition (identical or near-duplicate reformulations). Table 4 summarizes trajectory properties, and Table 5 reports their usage. We interpret trajectories with three stability measures: *Dense Similarity* is the cosine similarity between consecutive query embeddings (Qwen3-Embedding-0.6B [53]); *Jaccard Similarity* is the lexical overlap over lowercased, whitespace-tokenized word sets; and *Result Overlap* is the Jaccard overlap between retrieved document identifier sets for consecutive queries (definitions in Appendix B).

The "Drill-Down" Bias: Across intents, agents mainly tighten constraints via local edits or pivot across nearby facets, while explicit broadening is consistently the least-used move (under 11%; Table 5). This imbalance suggests that agents are more comfortable focusing on a local neighborhood of the query space rather than stepping back to relax constraints and reconsider alternatives. Exploration

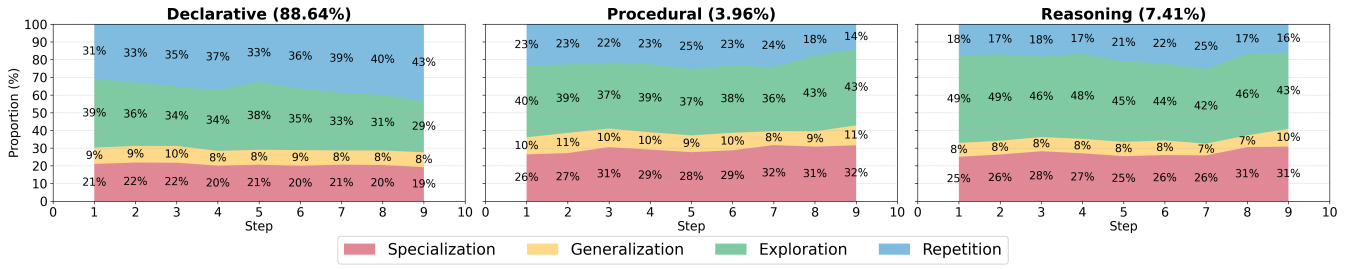


Figure 4: Step-wise trajectory distribution trends for the first 10 steps across different task intents. Each sub-figure illustrates the evolving proportions of Specialization, Generalization, Exploration, and Repetition, as the search session progresses.

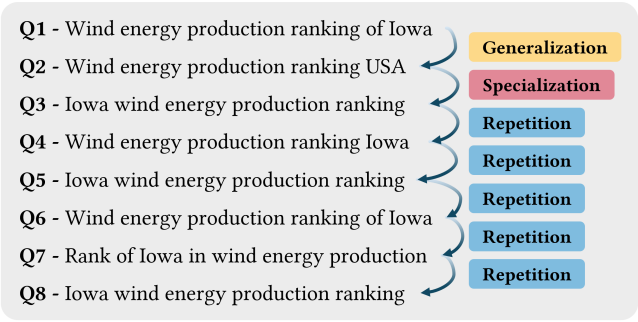


Figure 5: A Declarative retry-loop example dominated by near-duplicate reformulations.

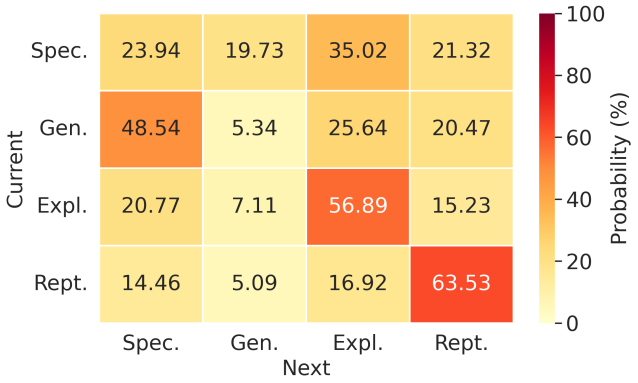


Figure 6: Trajectory transition matrix (row-wise normalized).

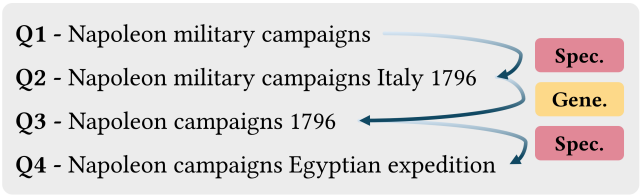


Figure 7: A reset-then-refine example: Specialization → Generalization → Specialization.

is common (roughly 36–48%), but pivots tend to induce larger evidence turnover and slower transitions, making them costlier than incremental refinement (Table 4). Together, these signatures suggest that agents often continue local edits in high-stability regimes rather than deliberately broadening and re-planning, making such regimes useful targets for future controller analysis.

Intent Differences: Although all intents share the same move vocabulary, they exhibit different interaction patterns (Table 5). Declarative sessions are most prone to retry-like behavior, with Repetition at about one-third of moves, consistent with agents re-issuing near-duplicate queries when evidence remains elusive. Reasoning sessions, in contrast, sustain the highest pivoting (Exploration near 48%) with lower retrying, suggesting a broader search over sub-questions. Procedural sessions more often combine pivots with subsequent constraint tightening, aligning with an "explore then refine" workflow in step-by-step tasks.

Stability as a Diagnostic: Table 4 reveals a stability spectrum. Repetition largely preserves retrieved results (Result Overlap ~78%), whereas Exploration induces major evidence turnover (Result Overlap ~7%). Specialization and Generalization fall between these extremes, typically preserving part of the retrieved context while steering the query. This suggests that sustained high-stability runs, especially Repetition, can serve as diagnostic markers of retry-like behavior, whereas Exploration and Specialization more often coincide with evidence turnover. Figure 5 illustrates the contrast in a real session. The agent briefly generalizes and then immediately re-specializes (Q1→Q2→Q3), and the interaction then transitions into a high-stability retry loop (Q3–Q8) where the intent remains largely unchanged despite minor wording edits. This example helps explain why Repetition is prominent in Declarative tasks and highlights an intervention point. Detecting such loops could support future tests of strategy-switching policies (e.g., to Exploration) to break the cycle. We connect these regimes to evidence traceability signals, such as term adoption, in Section 7.3.

Pacing Implications: Move types also differ in end-to-end pacing between requests (Table 4). Exploration is slower (median of 14.0s) than minor reformulations such as Repetition (6.0s), consistent with facet pivots inducing larger evidence turnover and higher processing cost. This makes strategy selection consequential. When pivots are expensive, agents may default to cheaper local edits, which motivates future tests of when local refinement should be replaced by a different move. Although explicit broadening is rare, it is often followed by re-specialization in the transition dynamics (Section 7.2), consistent with broadening acting as a brief reset rather than sustained re-planning.

Table 6: Mean CTAR under Aggregated vs. Last-step Evidence.

Trajectory type	Aggregated	Last-step
Specialization	78.35%	70.93%
Generalization	52.95%	49.44%
Exploration	69.59%	60.14%
Repetition	20.92%	19.77%
Overall	54.35%	48.54%

7.2 Temporal Dynamics of Search Strategies

We next study how query-reformulation strategies evolve over a session. Figure 4 traces the step-wise trajectory composition over the first 10 steps for each intent, and Figure 6 summarizes how moves transition from one step to the next.

Trends over Steps: Strategy use shifts over time. Early steps mix facet pivots, retries, and constraint adjustments, then diverge by task type (Figure 4). Declarative sessions gradually concentrate on retries, consistent with late-stage high-stability behavior; Procedural sessions maintain substantial pivoting but increasingly emphasize refinement; and Reasoning sessions sustain pivoting with consistently low retrying. We report the full step-wise proportions in Figure 4 and focus here on higher-level directional shifts.

Transition Mechanisms: Figure 6 helps explain these trends by showing which moves persist as runs and which act as resets. Exploration and Repetition often form multi-step runs, consistent with the step-wise patterns in Figure 4 where pivoting and retrying persist across consecutive steps. In contrast, broadening frequently acts as a brief reset. Nearly half of Generalization moves are followed by Specialization, suggesting that agents relax constraints momentarily before re-introducing them.

Case Study - The "Reset-then-Refine" Pattern: Figure 7 illustrates a reset-then-refine sequence. The agent specializes a broad topic by adding constraints (Napoleon campaigns → Italy 1796), then generalizes by removing them (a shorter, broader query), and re-specializes toward a different facet (Egyptian expedition). Query-length changes match our definitions, where Specialization tends to lengthen queries, while Generalization shortens them (Table 4). This is consistent with Generalization acting as lightweight backtracking to switch refinement branches rather than sustained broadening.

7.3 Context-Driven Term Adoption Rate (CTAR)

The previous sections show that agents frequently pivot, refine, and retry across steps, and that these strategies shift over a session's lifespan. A central question is whether such multi-step behavior is traceable to retrieved evidence. Since the logs do not reveal what an agent actually attends to or retains, we study a more limited signal. We test whether new query terms introduced at step k+1 appear in the evidence returned before that step. We use CTAR as defined in Section. 4.2 and report both last-step and aggregated variants.

Lexical Traceability and Recency. The CTAR analysis indicates that a substantial fraction of newly introduced query terms can be lexically traced to retrieved evidence. Table 6 reports the mean CTAR under aggregated versus last-step context.

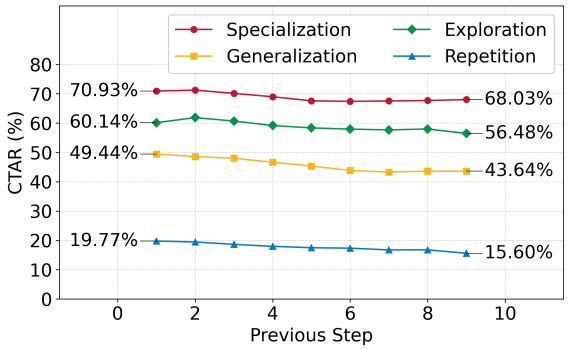


Figure 8: CTAR scores across previous steps.

Overall, more than half of newly introduced query terms are present in the aggregated evidence context, with a mean CTAR of 54.35%. Aggregated context adds 5.81 percentage points over last-step evidence, which is consistent with strong recency effects and some additional lexical traceability to earlier steps [27]. CTAR also varies substantially by trajectory type. Specialization and Exploration have much higher aggregated CTAR than Repetition, at 78.35% and 69.59% compared with 20.92%.

CTAR is intentionally a *lexical* traceability measure based on token overlap, rather than a semantic variant, to avoid embedding-model dependence and uninterpretable similarity-threshold choices. It is therefore best viewed as a conservative audit signal. A low CTAR score does not necessarily imply that the agent ignores evidence, since semantic paraphrases and abstractions are not counted. Conversely, a high CTAR score shows that new terms are explicitly present in the returned context, but it does not by itself establish that the evidence caused the reformulation.

Multi-step Context Traceability. We further examine how CTAR varies when tracing new terms against evidence from different historical steps within a session. Figure 8 summarizes the CTAR measured against progressively older evidence contexts. For example, a value of 70.93% for Spec at "previous step 1" means that, for Specialization transitions, 70.93% of newly introduced terms in q_{k+1} also appear in the evidence retrieved at the immediately preceding step E_k, corresponding to the Last Step context in Table 6. "Previous step 2" analogously traces against E_{k-1}, and so on. The curves show that new terms are also lexically traceable to older retrieved evidence. However, because retrieved evidence can overlap across steps, this analysis does not isolate the unique contribution of older evidence. We interpret Figure 8 as age-specific traceability, not as proof that the agent directly used earlier documents.

Taken together, these results show that query reformulation is often consistent with evidence-traceable term adoption. New query terms are frequently present in retrieved context, with the strongest signal from the most recent step and weaker but non-trivial traceability to earlier steps. This supports using CTAR as a lightweight diagnostic for cross-step evidence use, while leaving causal interpretation and semantic evidence use to future work.

8 Discussion, Implications, and Limitations

Our analysis provides an observational view of how agentic clients use a shared retrieval backend. The patterns below should be read

as diagnostic signals and hypotheses for future systems, rather than evidence that a behavior improves or harms downstream answer quality. The logs do not include success labels such as answer correctness, task completion, or user satisfaction, and they do not expose client-side prompts, memory, or control policies.

Repetition as a Candidate Stall Signal. In Declarative sessions, repetition increases to 42.68% by Step 9 (Figure 4). This pattern is consistent with near-duplicate loops, although the logs cannot determine whether repetition reflects stalled search, verification, cautious evidence checking, or client-side constraints. Repetition can therefore serve as a candidate signal for future controller policies. Such policies could test whether sustained lexical overlap should lead to a broader query, a different reformulation move, or human review [11, 43].

Intent-Adaptive Resource Allocation. Retrieval depth is largely rigid, with 91.64% of requests using $K \in \{1, 5, 10\}$, despite intent-dependent differences in observed usage, such as deeper retrieval for Procedural sessions than Declarative sessions. This suggests that many clients treat retrieval depth as a static parameter. Future architectures could evaluate intent-aware budgeting policies that adjust compute and retrieval depth across intents and steps, rather than relying on fixed K choices [19].

Evidence Grounding as an Audit Signal. The gain in CTAR from aggregated context (+5.81 pp over the last-step context) is consistent with cross-step lexical traceability, but it does not show that agents actively synthesize or attend to historical evidence. Similarly, the contrast between Specialization (78.35% CTAR) and Repetition (20.92% CTAR) shows that lower lexical evidence adoption co-occurs with retry-like transitions, not that low CTAR causes retries. CTAR can therefore be used as a lightweight audit signal for future context management modules that cache prior evidence and surface useful terms for later query formulation [45].

9 Conclusions

We study agentic search behavior *in the wild* through 14.44M DRGym requests [7], converting raw API logs into sessions and analyzing session-level structure, step-wise query transitions, and evidence traceability. Our central takeaway is that multi-step agentic search exhibits measurable intent-conditioned reformulation patterns, even when only API-level traces are available. The diversity analyses in Section 3 show that the logs are not dominated by a small set of repeated prompts or by the selected benchmark tasks, supporting their use for large-scale behavioral analysis. At the aggregate level, most multi-turn sessions remain short, while retrieval depth is often fixed within a session. This suggests that many current agentic clients rely on repeated query reformulation more than adaptive control of retrieval parameters.

The intent-conditioned and trajectory-level analyses further show that agentic search is not a uniform iterative process. Declarative sessions exhibit more retry-like behavior, while Procedural and Reasoning sessions show different mixtures of refinement, exploration, and repetition. Across trajectories, agents show a drill-down bias, favoring local refinement and facet pivots over deliberate broadening or backtracking. The transition patterns and case studies also show that high-stability runs can emerge within sessions, especially in the context of Declarative tasks. These patterns should not be read as direct evidence of agent success or failure, since the logs do not contain downstream outcome labels. They do, however, provide useful diagnostic structure for studying when an agent continues local edits, when it pivots to a new facet, and when it returns to near-duplicate queries.

To study evidence traceability without clicks or client-side attention signals, we introduce CTAR and measure whether newly introduced query terms appear in retrieved context from earlier steps. We find that many new query terms are lexically traceable to returned evidence, with the strongest signal from the most recent retrieval and weaker but non-trivial traceability to earlier steps. This finding does not establish that agents causally used or retained those documents, nor does it capture semantic paraphrases or abstractions. Instead, CTAR provides a lightweight lexical audit signal for whether query reformulation is consistent with evidence returned by the retrieval backend.

Taken together, the obtained results point to several directions for designing and evaluating agentic IR systems. The move distributions, transition dynamics, and case studies in Section 7 suggest that high-stability loops may be useful signals for future controller policies. Such policies could test when repeated local edits should be followed by broader queries, facet pivots, or adaptive retrieval budgeting. Our intent-conditioned analyses also indicate that a single global reformulation policy may be insufficient, since refinement, exploration, and retrying appear with different frequencies across task types. Finally, CTAR can inform future memory and context-management designs by exposing whether query terms remain traceable to prior evidence, while leaving causal evidence use and downstream utility to future evaluation.

Beyond releasing the DRGym-derived dataset and the analysis protocol, future work can connect reformulation moves to downstream answer quality, study which forms of backtracking and evidence reuse are beneficial, and evaluate whether controller interventions improve efficiency or answer quality. More broadly, we hope these measurements and case-driven diagnostics provide a foundation for intent-aware analysis and control of agentic search systems under reproducible retrieval settings.

Acknowledgments

The Portuguese researchers were supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by the Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (https://doi.org/10.54499/UID/50021/2025), and UID/PRR/50021/2025 (https://doi.org/10.54499/UID/PRR/50021/2025). João Coelho was additionally supported through the Ph.D. scholarship with reference PRT/BD/153683/2021, under the CMU Portugal Program.

A Log Sessionization Procedure

This section describes our log sessionization pipeline, including the semantic continuity model used to link adjacent queries and the per-IP online assignment rules used to form sessions.

Step 1: Train a semantic continuity model.
(1) Training pairs: We randomly sample ~200K queries and pair each query with the nearest-in-time query from the same IP.
(2) Pair labels: We label each pair as *same-session* vs. *different-session* using an LLM-as-a-judge prompt with `gpt-5-nano-2025-08-07` (Appendix D.1).
(3) Pair representation: We encode each query with Qwen3-Embedding-0.6B [53] and use the resulting embeddings to construct a fixed dense feature vector for each query pair as input to the downstream neural classifier.
(4) Classifier: We train a 3-layer MLP to output a continuity score in $[0, 1]$. The hidden-layer dimensions are 1024, 512, and 256, and the model achieves a held-out accuracy of 0.9419.

Step 2: Sessionize online per IP (with validation).
(1) Per-IP assignment: We process queries in order and maintain active sessions for each IP. For an incoming query q_t, we score it against each active session using the session's most recent query and assign q_t to the highest-scoring session if the score ≥ 0.5; otherwise we start a new session.
(2) Temporal hard cutoff: Beside the continuity score, if the gap to the candidate session's last query exceeds 10 minutes, we start a new session.
(3) Sanity check: We manually inspect 100 random sessions for coherence; after excluding four unusually long sessions, the remaining sessions are centered on a single objective.

B Auxiliary Metric Definitions

This section defines the auxiliary metrics used throughout our analyses and provides their formal notation and formulas for reproducibility (Table 7).

Notation: For a session $s = (q_1, \ldots, q_{|s|})$, let $\mathbf{v}_t$ denote the dense embedding of query q_t (and $\cos(\cdot, \cdot)$ the cosine similarity). W_t is the set of normalized tokens from q_t after lowercasing and stopword-aware tokenization $\mathrm{WS\_tok}(\cdot)$. D_t is the set of retrieved evidence returned for query q_t (at the logged retrieval depth).

Table 7: Summary of auxiliary metrics used in our analyses.

Metric	Formula
Initial–Final Gap	$\mathrm{Gap}(s) = 1 - \cos(\mathbf{v}_1, \mathbf{v}_{\lvert s\rvert})$
Dense Sim.	$\mathrm{DenseSim}(q_t, q_{t+1}) = \cos(\mathbf{v}_t, \mathbf{v}_{t+1})$
Jaccard Sim.	$\mathrm{Jac}(q_t, q_{t+1}) = \dfrac{\lvert W_t \cap W_{t+1}\rvert}{\lvert W_t \cup W_{t+1}\rvert}, \quad W_t = \mathrm{WS\_tok}(\mathrm{lower}(q_t))$
Result Overlap	$\mathrm{Overlap}(q_t, q_{t+1}) = \dfrac{\lvert D_t \cap D_{t+1}\rvert}{\lvert D_t \cup D_{t+1}\rvert}$

C Representative Query Examples

This section presents real representative query examples from our logs to help interpret our intent (Table 8) and trajectory (Table 9) labels.

Table 8: Representative Queries for Intent Categories.

Intent	Example Queries
Declarative	1. Who owns Handi-Snacks? 2. Definition of home food store 3. Ansel Adams residences in Yosemite National Park
Procedural	1. Best instructions for homemade reading shelf 2. Practical driving tips for beginners 3. How to change a constitution?
Reasoning	1. Why is gas so expensive? 2. Does advertising help or harm us? 3. Why are minority rights important?

Table 9: Representative Transitions Trajectories. Each entry illustrates a step-wise reformulation ($q_k \rightarrow q_{k+1}$).

Type	Step	Example Queries
Specialization	1	Recent climate data Durban
	2	Average temperature in Durban 2025
Generalization	1	Key events that ended Hitler's dictatorship
	2	Hitler's dictatorship
Exploration	1	Headquarters of Oberoi Hotels
	2	Parent company of Oberoi Hotels
Repetition	1	Handi-Snacks parent company
	2	Who owns Handi-Snacks

D LLM-as-a-judge Prompts and Parsing Details

This section provides the exact LLM-as-a-judge prompts for reproducibility.

D.1 Query-pair Continuity Judgment Prompt

```
[SYSTEM]
You must label query pairs for a DeepResearch search agent. The agent fans out
      several queries to answer ONE user question.
Answer YES if both queries would naturally be used for the same research task or
       user question (same core topic), even if they cover different aspects or
       levels of detail.
Answer NO if the queries clearly correspond to different questions, even if they
       share broad words like 'WWII', 'health', or 'economy'.

[USER]
Query 1: <<query1>>
Query 2: <<query2>>
For a DeepResearch agent that fans out queries to answer ONE user question,
would these two queries belong to the same research task?
Answer YES or NO only.
```

D.2 Session-level Intent Classification Prompt

```
[SYSTEM]
You are an expert search intent classifier.

[USER]
"Session Queries:\
<<joined_queries>>
Classify the user intent of this session into exactly ONE of these three
      categories:
1. Declarative: Asking for simple facts, definitions, entity attributes, or
      lists (e.g., 'who is', 'what is', 'release date').
2. Procedural: Asking for steps, methods, tutorials, or guides (e.g., 'how to',
      'guide for', 'fix error').
3. Reasoning: Asking for comparisons, planning, analysis, multi-hop reasoning,
      or creative generation (e.g., 'difference between', 'best plan for', 'why
       is').
Output ONLY the category name (Declarative, Procedural, or Reasoning).
```

D.3 Step-wise Trajectory Classification Prompt

```
[SYSTEM]
You are an expert search behavior analyst.

[USER]
Query 1 (Previous): <<PLACEHOLDER: q_k>>
Query 2 (Current):  <<PLACEHOLDER: q_{k+1}>>
Analyze the search behavior evolution from Query 1 to Query 2 for an autonomous
      agent.
Classify the transition into exactly ONE of these four categories:
1. Specialization (Vertical Deepening): Query 2 is MORE specific than Query 1 by
       adding constraints/details (q2 \subset q1). (e.g., 'apple' -> 'green
      apple nutritional value').
2. Generalization (Vertical Broadening): Query 2 is MORE general than Query 1 by
       removing constraints/abstracting (q2 \supset q1). (e.g., 'green apple
      nutritional value' -> 'benefits of fruits').
3. Exploration (Horizontal Expansion within the same domain/task): Query 2 is
      NOT simply more specific or more general. It shifts to a different aspect
      /subtopic/related entity but still within the same overall topic/domain.
      (e.g., 'green apple nutritional value' -> 'green apple recipes' or 'MRI
      Scans' -> 'CT Scans').
4. Repetition (Stationary): Query 2 is semantically equivalent to Query 1. It is
       a paraphrase, reformatting, or synonym replacement with NO significant
      change in intent (e.g., 'green apple value' -> 'nutritional value of
      green apple').
Output ONLY the category name (Specialization, Generalization, Exploration, or
      Repetition).
```

References

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *International Conference on Research and Development in Information Retrieval (SIGIR)*.

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations (ICLR)*.

[3] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. 2011. Query reformulation mining: models, patterns, and applications. *Information Retrieval*.

[4] Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum*.

[5] Aaron Brown and Matt Saner. 2025. The Agentic AI Security Scoping Matrix: A framework for securing autonomous AI systems. AWS Security Blog. Published: 21 Nov 2025. Accessed: 29 Dec 2025. (2025). https://aws.amazon.com/cn/blogs/security/the-agentic-ai-security-scoping-matrix-a-framework-for-securing-autonomous-ai-systems/.

[6] Wei-Lin Chiang et al. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. (2024). arXiv: 2403.04132.

[7] João Coelho et al. 2025. DeepResearchGym: A Free, Transparent, and Reproducible Evaluation Sandbox for Deep Research. (2025). arXiv: 2505.19253.

[8] Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, and Jaime Teevan. 2014. Understanding User Behavior Through Log Data and Analysis. *Ways of Knowing in HCI*.

[9] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An Eye-Tracking Study of Query Reformulation. In *International Conference on Research and Development in Information Retrieval (SIGIR)*.

[10] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *International Conference on Web Search and Data Mining (WSDM)*.

[11] Henry A. Feild, James Allan, and Rosie Jones. 2010. Predicting Searcher Frustration. In *International Conference on Research and Development in Information Retrieval (SIGIR)*.

[12] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*.

[13] Pedro Gomes, Bruno Martins, and Luís Cruz. 2019. Segmenting User Sessions in Search Engine Query Logs Leveraging Word Embeddings. In *International Conference on Theory and Practice of Digital Libraries (TPDL)*.

[14] Google. 2025. Gemini 3 Developer Guide (model id: gemini-3-flash-preview). Google AI for Developers Documentation. Gemini 3 models in preview; model IDs listed in documentation. (2025). Retrieved Jan. 18, 2026 from https://ai.google.dev/gemini-api/docs/gemini-3.

[15] Kunal Handa et al. 2025. Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. (2025). arXiv: 2503.04761.

[16] Jeff Huang and Efthimis N. Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Conference on Information and Knowledge Management (CIKM)*.

[17] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. 1998. Real life information retrieval: a study of user queries on the web. *SIGIR Forum*.

[18] Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. 2019. DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

[19] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[20] Jiahe Jin, Abhijay Paladugu, and Chenyan Xiong. 2025. Beneficial Reasoning Behaviors in Agentic Search and Effective Post-training to Obtain Them. (2025). arXiv: 2510.06534.

[21] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *International Conference on Research and Development in Information Retrieval (SIGIR)*.

[22] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Conference on Information and Knowledge Management (CIKM)*.

[23] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. (2020). arXiv: 2004.04906.

[24] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generatio. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

[25] Ido Levy, Ben wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2025. ST-WebAgentBench: A Benchmark for Evaluating Safety and Trustworthiness in Web Agents. In *ICML Workshop on Computer Use Agents (ICML WCUA)*.

[26] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. (2024). arXiv: 2412.05579.

[27] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*.

[28] Xiao Liu et al. 2023. AgentBench: Evaluating LLMs as Agents. (2023). arXiv: 2308.03688.

[29] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Communications of the ACM*.

[30] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: a benchmark for General AI Assistants. (2023). arXiv: 2311.12983.

[31] Reiichiro Nakano et al. 2022. WebGPT: Browser-assisted question-answering with human feedback. (2022). arXiv: 2112.09332.

[32] Lunyiu Nie, Nedim Lipka, Ryan A. Rossi, and Swarat Chaudhuri. 2025. FlashResearch: Real-time Agent Orchestration for Efficient Deep Research. (2025). arXiv: 2510.05145.

[33] OpenAI. 2025. GPT-5 nano Model. OpenAI API Documentation. Accessed: 2025-12-29. (2025). https://platform.openai.com/docs/models/gpt-5-nano.

[34] OpenAI. 2025. How People Use ChatGPT. Tech. rep. OpenAI.

[35] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Visual and Semantic Information. (2022). arXiv: 2211.15848.

[36] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. (2024). arXiv: 2406.17557.

[37] Long Phan, Alice Gatti, Ziwen Han, et al. 2025. Humanity's Last Exam. (2025). arXiv: 2501.14249.

[38] Yujia Qin et al. 2023. ToolLLM: facilitating large language models to master 16,000+ real-world APIs. (2023). arXiv: 2307.16789.

[39] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: a review of current perspectives and future directions. *Journal of Information Science*.

[40] Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *International Conference on World Wide Web (WWW)*.

[41] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. (2023). arXiv: 2302.04761.

[42] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*.

[43] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. 2007. Information re-retrieval: repeat queries in yahoo's logs. In *International Conference on Research and Development in Information Retrieval (SIGIR)*.

[44] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Conference on Human Factors in Computing Systems (CHI)*.

[45] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

[46] Kelsey Urgo and Jaime Arguello. 2022. Learning assessments in search-as-learning: a survey of prior work and opportunities for future research. *Information Processing and Management*.

[47] Zhefan Wang, Ning Geng, Zhiqiang Guo, Weizhi Ma, and Min Zhang. 2025. Human vs. Agent in Task-Oriented Conversations. (2025). arXiv: 2509.17619.

[48] Zora Zhiruo Wang, Yijia Shao, Omar Shaikh, Daniel Fried, Graham Neubig, and Diyi Yang. 2025. How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations. (2025). arXiv: 2510.22780.

[49] Ryen W. White and Steven M. Drucker. 2007. Investigating behavioral variability in web search. In *International Conference on World Wide Web (WWW)*.

[50] Jialong Wu et al. 2025. WebWalker: Benchmarking LLMs in Web Traversa. In *Annual Meeting of the Association for Computational Linguistics*.

[51] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

[52] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. (2023). arXiv: 2210.03629.

[53] Yanzhao Zhang et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. (2025). arXiv: 2506.05176.
[54] Yilun Zhao et al. 2025. SciArena: An Open Evaluation Platform for Foundation Models in Scientific Literature Tasks. (2025). arXiv: 2507.01001.
[55] Lianmin Zheng et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *International Conference on Neural Information Processing Systems (NeurIPS)*.
[56] Lianmin Zheng et al. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *International Conference on Learning Representations (ICLR)*.
[57] Jianan Zhou, Fleur Corbett, Joori Byun, Talya Porat, and Nejra van Zalk. 2025. Psychological and behavioural responses in human-agent vs. human-human interactions: a systematic review and meta-analysis. (2025). arXiv: 2509.21542.
[58] Shuyan Zhou et al. 2023. Webarena: a realistic web environment for building autonomous agents. (2023). arXiv: 2307.13854.