PRIVaThe: An Annotated Dataset of Multi-Objectives Web Search Sessions

Claire Ibarboure, Ludovic Tanguy, Franck Amadieu, Josiane Mothe
Univ Toulouse, Université Toulouse-Jean Jaurès, CNRS, CLLE, Toulouse, France
firstname.lastname@univ-tlse2.fr

Abstract

This paper presents PRIVaThe, a new French-language dataset, consisting of 200 web search sessions from 100 participants performing two multi-objective, multi-hop tasks, designed to enable cross-user comparison of session-level search strategies. Unlike existing datasets that capture only query sequences or final answers, PRIVaThe provides explicit sub-objective decomposition traces for each session. We automatically annotate 3,162 queries with their addressed sub-objective(s) using validated open-weight LLMs (Mistral, LLama3, and Gemma) against human gold annotations. This annotation enables systematic analyses of how users distribute and sequence sub-objectives throughout their sessions, revealing distinct search strategies such as logical, global, and exploratory approaches.

Keywords: Dataset, Search strategies, Automatic annotation, French language, Multi-hop search, Complex information need

1. Introduction

Complex information search tasks are common in real-world search scenarios. When planning a vacation, comparing job offers, or researching a medical condition, users must decompose high-level goals into manageable sub-problems, execute multiple related searches, and synthesise information across diverse sources. Previous work shows that users' queries can vary substantially even for simple search tasks: Alaofi et al. (2022) for instance, observed 19 different first query formulations for the question "How do you cook Beef Stroganoff for a dinner party?" among 108 participants. A natural question is whether such variability persists when tasks become complex and require decomposition into sub-problems, or whether users instead converge on similar search strategies. Understanding how humans strategically structure these multi-hop searches —not only which queries they issue, but also how they decompose goals and integrate findings— is essential for studying human cognition and can inform the design of search engine. Addressing this question requires collecting comparable search sessions, i.e. multiple users solving the same complex task, so that differences in strategy can be analysed under a shared information need.

Existing datasets fall short of capturing these critical dimensions of search behaviour. Large-scale query log datasets provide real-world search patterns, but lack crucial context about users' underlying goals and decomposition strategies (Wang and Zhai, 2007; Kacprzak et al., 2019; Sharifpour et al., 2023). Question-answering datasets like HotpotQA demonstrate the need for multi-hop reasoning but provide only final answers without capturing the intermediate search process (Yang et al., 2018; Mavi et al., 2024). Annotated datasets like CoST, which tags local query reformulation tactics, offer only query-to-query transitions rather than session-level strategic patterns (Dosso et al., 2021). This gap is particularly acute for under-resourced languages like French, where comprehensive search session datasets with semantic annotations do not exist.

We release PRIVaThe[1] (*Parcours de Recherche d'Information avec Variations Thématiques - Information Retrieval Paths with Thematic Variations*), a French dataset designed to study session-level strategies in complex, multi-objective Web search tasks. Specifically, we aim to analyse how users decompose a goal into sub-objectives and how they sequence and interleave those sub-objectives over time, using queries as main behavioural traces. We prioritise depth per task over breadth across tasks: our analyses require many sessions solving the same complex information need in order to identify and robustly compare session-level strategies. We therefore include two structurally similar but topically distinct tasks to test generality while keeping the design controlled.

PRIVaThe enables the analysis of both the search process and the underlying cognitive strategies employed in complex information search tasks. More precisely, with this dataset, we can study strategies related to planning (i.e. finding a solution to a problem by proposing, if necessary, a segmentation into sub-objectives (Sharit et al., 2008)) through the formulation of the queries. As a result, it is possible to use an inductive approach to identify a typology of the different strategies used when users have to respond to a complex task involving sub-objectives relating to different topics.

PRIVaThe consists of 200 search sessions from

[1] https://gitlab.huma-num.fr/clle/privathe

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7380–7390
11–16 May 2026.

100 participants completing two carefully designed multi-objective search tasks, yielding a total of 3,162 queries freely formulated by participants and submitted to the Google search engine. Each task was constructed to include two predefined sub-objectives (A and B) drawn from distinct topical domains, ensuring that participants had to coordinate searches across unrelated information spaces. We collected time-stamped search histories -including both search engine result pages (SERPs) and websites as well as the content of the SERPs.

To analyse how users strategically structure such complex searches, we annotated each query according to which sub-objective(s) it addresses. For example, does the query target animated films, operating systems, both simultaneously, or neither? This multi-label annotation enables systematic analysis of how users navigate between sub-objectives, when they formulate queries spanning multiple objectives, and what overall strategic patterns emerge across sessions. All queries are automatically annotated using open-weight LLMs (Mistral, Llama3, Gemma) after validating their annotation capabilities against human gold standards (achieving F1-Scores exceeding 0.89).

Our main contributions are:
(1) PRIVaThe dataset, featuring search sessions with explicit goal decomposition traces corresponding to a total of 3,162 queries from 100 participants on 2 multi-objective information search tasks,
(2) a validated LLM-based annotation methodology for identifying sub-objectives in queries using open models,
(3) an analysis showing evidence of distinct search strategies (logical, global, exploratory) based on sub-objective sequencing patterns,
(4) a comprehensive data release that includes timestamped queries, visited pages (SERPs and websites), user demographics, and domain knowledge assessments enabling diverse research applications.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the methodology for designing complex information search tasks. Section 4 presents the LLM-based annotation. Section 5 details the dataset, Section 6 reports initial analyses, and Section 7 concludes.

2. Related work

2.1. Datasets

Evaluation forums provide numerous data collections focused on document relevance to queries across different languages, e.g. TREC[2] with MS-Marco in English (Nguyen et al., 2016). Fewer datasets capture complex search tasks. HotpotQA, a multi-hop question answering dataset (Yang et al., 2018) shows the need for reasoning across multiple documents, but provides only the final answer without capturing the intermediate search process. The TREC Conversational Assistance Track (CAsT) (Owoicho et al., 2022) defines topics each associated with a user profile and multi-turn conversations, but integrates the generation part, thus not focusing on the search process itself. Search log datasets offer real-world query sequences, but lack crucial context about users' decomposition strategies (Wang and Zhai, 2007; Kacprzak et al., 2019; Sharifpour et al., 2023). In CoST, a French experimental dataset, reformulation tactics (the transition from query Q_n to query Q_{n+1} within the same information need) are annotated according to whether they reflect an *exploration* or *exploitation* of the current research object for research tasks of varying complexity (Dosso et al., 2021). This annotation provides only a local indication without offering an overall representation of behaviour at the level of the search session.

There is a fundamental limitation in existing datasets: we can observe what users search for, but not how they strategically decompose complex tasks. The PRIVaThe dataset we developed fills this gap, which is increasingly critical as search systems incorporate agentic capabilities that may require models of human strategic reasoning (Hughes et al., 2025). Moreover, this dataset contributes to French, an under-resourced language for search session analysis.

2.2. Annotation by LLMs

Linguistic annotation of data is necessary for analysing phenomena such as behavioural patterns. Manual annotation is costly and modern generic pretrained LLMs are now usable to provide complex annotation such as pragmatic annotation and discourse analysis (Yu et al., 2024) or semantic sentence structure analysis (Ettinger et al., 2023). Multilingual data has also accurately been annotated in such a way (Nemkova et al., 2025).

Throughout the literature, we observe that LLMs such as GPT-4 are sometimes as effective as human annotators on standard annotation tasks (Nemkova et al., 2025). As fine-tuning is both costly and most of the time unfeasible due to the lack of available data, the prompt is paramount and requires a cautious design. Zero-shot prompts (without prior annotation examples) and few-shot prompts (where annotation examples are provided upstream) have been reported as effective for text annotation (Reynolds and McDonell, 2021; Nemkova et al., 2025). Better performance is also observed when the language of the prompt is aligned with that of the corpus (Nemkova et al.,

[2]TREC - Text REtrieval Conference www.nist.gov

2025). We ourselves tested different prompt formulations to obtain the annotation that resembled the human annotation as closely as possible.

Automatic annotation with LLMs can be done without fine-tuning. However, the use of LLMs varies depending on the type of annotation or the data to be annotated, so it is necessary to experiment in order to obtain satisfactory annotation.

3. Dataset Construction Methodology

This section presents the methodology and describes the collected data, including qualitative and statistical summaries.

3.1. Complex Information Search: Definition and Framework

We consider a search session to be *complex* when the user needs to formulate multiple queries to satisfy their information need. When defining the tasks participants will perform, we kept in mind our goal to collect sessions (a) consisting of distinct information search sub-objectives (b) requiring multiple interaction hops, as we will see in this section.

Sub-Objectives In our framework, we operationalise complex information search tasks by decomposing them into sub-objectives. Sub-objectives correspond to distinct topical components that must each be resolved to complete the overall task. Each sub-objective targets a separate search space (e.g. animated films, operating systems), requiring users to formulate queries across multiple topics. We deliberately design sub-objectives with clear topical boundaries to enable deep analysis of users' search strategies such as the identification of sub-objective/domain shifts in user's queries. This task design allows analyses on how users navigate between sub-objectives, whether queries involve one or several sub-objectives at a time, etc. Such analyses could help understanding how users organise their search strategies when faced with complex information needs. For a detailed study of these behavioural variations within a search, the number of domains should be limited. We limit ourselves to two main sub-objectives per search task.

Session Length To study behavioural variation between users, it is preferable to have long search sessions. For this reason, we designed tasks for which users likely need to visit many web pages, in addition to writing many queries. We choose tasks which topics are very common and widely described on the Web, since the complexity we study is not about finding rare information but rather on multi-hop searches. We formulated the instructions in order to encourage users to formulate several queries, whether to check conditions and/or to search for detailed information (see Section 3.2).

Additionally, we aimed to make tasks attractive despite their complexity to induce user involvement. We designed these tasks as engaging puzzles to stimulate interest while imposing a 20-minute time limit per task. This time constraint both reflects typical user time limitations and restricts time spent reading web pages to encourage query formulation. Moreover, we limited the duration of the tasks to avoid having an overly long experimental protocol that could tire participants and overload them cognitively.

3.2. Complex Task Design

To systematically compare behavioural variation, we controlled for task formulation by assigning identical information needs to all participants. This design ensures that differences in observed behaviour reflect variation in search strategy rather than differences in task interpretation. Furthermore, we introduced two distinct search tasks (see Figure 1) and deployed them across a large participant sample to enhance the robustness and generalisability of our findings.

Both tasks are conceived in the same way: their resolution required the users to pursue two distinct sub-objectives in order to establish a specific link between two clearly differentiated topics.

The first task (TSF for "Task System and Film") is to find the link between two apparently disconnected entities (see Figure 1). The two sub-objective topics are *animated film* and *operating systems*, and the overall targeted answer is the fact that Debian version names are based on the characters from the *Toy Story* series of animated films by Pixar (e.g. Debian 1.2 is named "Rex"). We formulated the task instructions so that the user must identify the animated film based on the names of its director and French dubbing actor only, and find the operating system based on its logos.

For the second task (TMR for "Task Mythology and Renaissance") the target information is a list of Italian Renaissance paintings depicting scenes of seduction of mortal women by the god Zeus in animal form (e.g. Leonardo's "Leda and the Swan"). More precisely, the user is required to identify three different paintings/events, indicating for each one the painter, animal, and seduced woman. Note that, to our knowledge and after a thorough research of our own, only two possible answers exist (although partial propositions can be considered). We also decided to identify Zeus as "Heracles' father" in the instructions in order to induce more search actions. For this second task, the two search spaces are *Greek mythology* and *Italian Renaissance*.

Figure 1 contains the English translation of the complete instructions given to the participants for both tasks (the original French version is part of delivered data collection).

Task 1 (TSF): What is the connection between the **animated films** directed by Pete Docter in which Henry Guybet lends his voice, and version 1.2 of the **operating system** represented by a spiral, associated with the one featuring the penguin?

- - - - - - - - - -

Task 2 (TMR): According to **Greek mythology**, Heracles' father transformed himself into an animal three times in order to seduce mortal (human) women. These transformations were the subject of numerous **Italian Renaissance** paintings.
To complete this task, you must fill in the following table, specifying for each of the three seductions the animal used in the transformation, the woman seduced, the title of the painting depicting the seduction, and the painter of the painting.
The paintings may be produced by different painters; however, across the three works, there must be three distinct animals and three distinct women represented.

Animal	Woman	Painting	Painter

Figure 1: Instructions given to participants for both tasks (translated from the French original version)

For 50 participants, we presented the version of the statement shown in Figure 1. For the other half of the participants, we presented a version that reversed the order of the two sub-objectives, to account the influence of the statement on the first query formulation.

3.3. Participants and Recruitment

We select adult native French speakers as participants to avoid comprehension and communication bias during the tasks. Participants were recruited on the authors' university campus through advertisements (flyers and email announcements) as well as in-class recruitment. The funding obtained after data collection began facilitated participant recruitment. Consequently, just under half of the participants received financial compensation as a €10 gift card. In addition, second-year undergraduate students received a small bonus point to their grade for one of their courses.

All 100 participants are students (ranging from bachelor's to doctoral level) or post-doctoral researchers in humanities and social sciences. 75% of users are aged between 18 and 25. We have a predominantly female sample with 75 women, 21 men, 3 non-binary individuals, and 1 person who did not wish to disclose their gender. Details are provided in the dataset material.

3.4. Data Collection Protocol

3.4.1. Experiment Procedure

First and foremost, the entire procedure for our experiment was validated by an Ethics Review Committee (ERC - agreement n°2023-614). The experiment begins with the signing of a consent form explaining the procedure of the experiment, as well as the agreements on the storage and dissemination of data, and all participants agreed. Participants then fill a short sociodemographic questionnaire.

After that, participants complete a simple test task which demonstrates what is expected in the rest of the experiment. In addition, before completing each of the two main tasks, participants answer a multiple-choice self-assessment questionnaire related to the areas covered by the task to determine their prior knowledge (the questionnaires are available in the delivered data collection[3]). They have a maximum of 20 minutes to complete each task once they have received the paper version of the instructions. They may abandon the task at any time or stop after writing down their answer, even if the time limit has not been reached.

The collected information –sociodemographic and the scores of self-assessment questionnaires– as well as the terms on data usage they agreed on are part of the dataset.

3.4.2. Material

We limited artificial constraints to maximise the ecological validity of the framework. For technical collection reasons discussed later in this section, the participants were nevertheless required to use the Google Chrome browser, but they were free to use it as they wished, e.g. open multiple tabs or use keyboard shortcuts. They were also allowed to formulate their queries in the language of their choice. However, participants were not allowed to use generative LLMs to search for information, but only traditional search engines such as Google. Indeed, we are interested in human strategies for filling an information need. We therefore do not want participants to be assisted in their work, particularly in reasoning and research, by generative LLMs. Moreover, participants conducted their search sessions on identical standard computers exclusively dedicated to the experiment in order to protect their anonymity.

The complete search and navigation history of the session was recorded using the *Export Chrome*

[3] https://gitlab.huma-num.fr/clle/privathe

order	id	date	time	title	url	visitCount	typedCount	transition
0	103	3/10/2023	10:48:43	henri guybet toy story - Recherche Google	https://www.googl	2	0	link
1	103	3/10/2023	10:48:42	henri guybet toy story - Recherche Google	https://www.googl	2	0	form_submit
2	102	3/10/2023	10:48:21	pete docter toy story - Recherche Google	https://www.googl	2	0	link
3	102	3/10/2023	10:48:20	pete docter toy story - Recherche Google	https://www.googl	2	0	form_submit
4	101	3/10/2023	10:47:52	monstre & co Rex - Recherche Google	https://www.google.c	2	0	link
5	101	3/10/2023	10:47:51	monstre & co Rex - Recherche Google	https://www.google.c	2	0	generated
6	100	3/10/2023	10:47:30	Historique des versions de Debian — Wikipé	https://fr.wikipedia.o	1	0	link
7	99	3/10/2023	10:47:22	version 1.2 debian linux - Recherche Google	https://www.google	2	0	link

Figure 2: Extract from one user's history recorded from Export Chrome History

History[4] extension. An extract of a history is shown in Figure 2, displaying all the titles of visited documents (SERPs and web pages) along with their timestamps. In addition, this history provides several pieces of information : the page ID (*id* - the last action appears first), the date on which the user accessed the page (*date*), the time at which the user -or participant- accessed the page (*time*), the title of the page (*title*), the URL of the page (*url*), the number of times the user accessed the page (*visitCount*), the number of times the user accessed the page by entering the address (*typedCount*) and the type of transition used to access the page (*transition*), i.e. the user's navigation path to the page (e.g. *link* means that the user followed a hypertext link located on another page).

We also extracted all SERPs and web pages opened by the user using the *ChromeCacheView*[5] tool, in order to obtain the content of the pages at the time of their visit.

3.5. Dataset Overview and Statistics

3.5.1. Descriptive statistics

In total, we collected 3,162 queries, comprising 1,518 for TSF and 1,644 for TMR.

To calculate the size of the sessions and analyse the queries, we extracted queries from the SERP titles in the history (e.g. line with order 0 `henri guybet toy story` in Figure 2). We removed identical consecutive queries, since SERPs are duplicated, for example, when users submit their query (see Figure 2 line with order 1 transition: form_submit) which causes a results page to open (see Figure 2 line with order 0 transition: link). On average, users formulate approximately 15 queries per session for TSF and 16 for TMR. Some users submitted a very large number of queries, up to 47 queries for TSF and 34 for TMR (see Figure 3 (a)). Users' queries are 4.64 words long on average for TSF, and 4.14 words for TMR (see details on Figure 3 (b)).

[4] https://chromewebstore.google.com/detail/export-chrome-history/dihloblpkeiddiaojbagoecedbfpifdj

[5] https://www.nirsoft.net/utils/chrome_cache_view.html

Based on the session history, we also determined the total duration of the search session, calculated as the difference between the first and the last action in the history for each task. Users visit more web pages (clicks) on average for TSF (14.63) than for TMR (6.86) (see Figure 3 (c)) and take longer to respond for TMR (15.47 minutes on average) than for TSF (9.95 minutes on average).

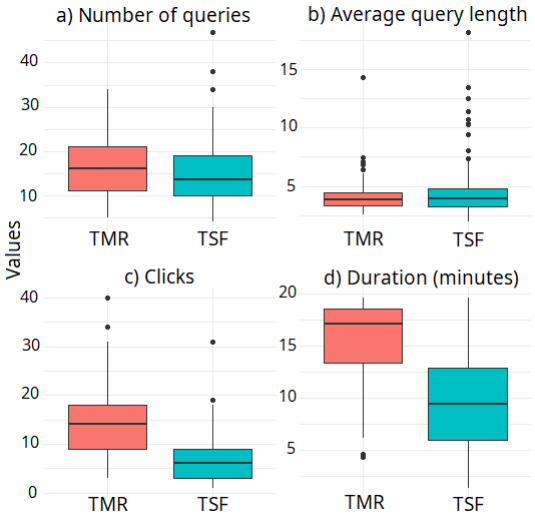

Figure 3: Main statistics: number of queries, query length, number of visited web pages, duration of sessions, for each task

3.5.2. Qualitative Description

Queries in PRIVaThe dataset take several forms, including natural language queries that follow a grammatical structure (e.g. questions such as `Pourquoi Zeus se transforme en aigle ?` *en: Why does Zeus transform into an eagle?*). There are sequences of keywords with no explicit syntagmatic link (e.g. `zeus aigle tableau` *en: Zeus eagle painting*). We also observe the use of querying operators such as + and quotation marks.

Some queries contain spelling or typing errors, which is typical in search engine logs (Hargittai, 2006). Most queries are in French, but there are also some queries in English (e.g. *bird hera zeus*

painting) or in both languages (code-switching).

The dataset is characterised by the presence of various proper names, which are essential for completing the search tasks. These include mythological and animated characters (e.g. Zeus, Heracles, Rex, etc.) as well as historical figures and contemporary people (e.g. Titian, Pete Docter, Ian Murdock, etc.), titles of paintings and films (e.g. The Rape of Europa, Toy Story, Monsters, Inc., etc.), and operating systems (e.g. Linux, Debian, Windows).

4. LLM-based Session Annotation

To examine behavioural variation in sub-objective formulation, we annotated the queries. A subset was manually labelled to establish a gold standard and annotation protocol. We subsequently assessed the extent to which open-weight, pre-trained large language models (LLMs) could replicate this human annotation framework.

4.1. Manual Annotation & Gold Standard

For manual annotation, we selected 100 random queries from different search sessions for both tasks. Two separate annotators (co-authors of this paper) annotated the queries according to the presence of search sub-objectives. Annotators were thus familiar with the tasks and knew the expected answers. They performed a multi-label classification using sub-objectives '1' and '2' codification and an 'Other' category if the query refers to elements other than those presented in the sub-objective definitions:

> **TMR - 1** Greek mythology: all the characters from mythology, as well as mythological animals and creatures.
>
> **TMR - 2** Art/Italian Renaissance: the entire lexical field related to art (painter, painting title) or the Italian Renaissance (dates, locations).
>
> **TSF - 1** Animated films: lexical field related to animated films (title, actor, director, voice, studio, character...)
>
> **TSF - 2** Operating systems: lexical field of operating systems (name, version)

The obtained Cohen's kappa values (0.94–1.00) indicate very high inter-annotator agreement (see Table 1). The few disagreements observed were primarily due to acknowledged annotation errors, which were easily resolved to produce a final adjudicated subset. The category 'Other' was used only once by a single annotator for the TMR query `La genèse` (*en: Genesis)*.

Cohen's Kappa	**AF**	**OS**
TSF	0.98	0.96
	GM	**IR**
TMR	1	0.94

Table 1: Cohen's Kappa scores for multi-label annotation of 100 queries by two judges. *AF = Animated film, OS = Operating system, GM = Greek mythology, IR = Italian Renaissance.*

4.2. LLM as an Annotator

Models To automatically annotate queries, we evaluated three small open weight LLMs, consisting of 7 to 8 billion parameters, which we ran locally on a small GPU, through the external Ollama platform[6]. We selected Llama 3 (8B) (Llama Team and al., 2024), Mistral 0.3 (7B) (Jiang et al., 2023) and Gemma 1.1 (7B) (Gemma Team and al., 2024).

Prompt In order to obtain the best annotation, we tested different prompts composed of at least the same information provided to human annotators, namely: the task statement, with an example answer and the definitions of the categories as described above.

In the prompt version yielding best results, we begin in English with a general context explaining the LLM's role (linguistic expert) and the annotation task (identifying the semantic categories of queries). We then introduce the query context by presenting the search task the users had to complete and an example of a response, both given in French. We then define the sub-objectives and the annotation guidelines in English. Finally, we present an example of expected output by proposing two annotated queries in French (see Appendix, also included in the resource). A fully French prompt yielded poorer annotations.

Method We annotated all queries separately so that adjacent queries did not influence annotation. We ran the prompt five times on the entire manually annotated dataset to evaluate annotation model consistency. We kept the default temperature suggested by Ollama (0.8), so output may vary from one run to another.

In some runs, we noticed errors preventing annotation identifying (e.g., output "Aigle, yes" instead of "A: yes"). In such cases, we assigned NA to the annotation.

Validation and Quality Assessment In order to evaluate the automatic annotation, we calculated the F1-scores for each sub-objective between the gold standard and the LLM annotation for the

[6] https://ollama.com/

different runs. Table 2 summarises the average F1-scores calculated for the 5 runs, for the three models, and for the two search tasks. As we can see, Mistral achieves the best results for TSF and Llama3 for TMR.

Sub-objectives	Mistral	Llama3	Gemma
AF	**0.94**	0.86	0.93
OS	**0.98**	0.95	0.97
GM	0.96	**0.97**	0.96
IR	0.89	**0.90**	0.85

Note. AF = Animated film, OS = Operating system, GM = Greek mythology, IR = Italian Renaissance.

Table 2: Average F1-Scores obtained by the models for TSF and TMR

The category "Other" is over-represented by LLMs compared to manual annotation, with an average of 22 times for Mistral for TSF and 14 times for Llama3 for TMR. We tried to limit the number of instances in this category by varying the prompt, but without success. We therefore decided to annotate all TSF queries with Mistral and all TMR queries with Llama3. To do this, we ran five more runs and used a majority vote for the final data.

For the final assessment of the resulting data, we obtain an F1-Score of 0.95 for the Animated Film category and 0.98 for the Operating System category with Mistral. For TMR, we obtain an F1-score of 0.98 for the Greek Mythology category and 0.88 for the Italian Renaissance category with LLama3. For the final annotation of the dataset, we found that some queries were annotated in the "Other" category only, even though they belonged to one of the two categories (e.g. `windows` = operating system category, `La-haut` = animated film category (French title for the movie Up)). Looking at other annotations, we saw other errors that we could not explain such a query that is not annotated in the category Greek mythology even though there is a name of the goddess Persephone (e.g. `persephone tableau renaissance`). On the other hand, some errors can be explained by typos in queries, for example, the query `montre inc link with linux 1.2` could not be categorised as Animated film because the title of the movie Monster Inc is misspelled (*montre* (en. *watch* or *show*) here is a misspelling of *monstre*). Moreover, some errors may result from translation errors (e.g. "monster and cie" instead of Monster Inc.).

In total, for the TMR subset, 799 queries are annotated as belonging exclusively to the 'Greek mythology' category, and only 80 to 'Art/Italian Renaissance', 742 in both categories at the same time, and 23 queries are annotated only in 'Other' category. For the TSF subset, 510 queries are annotated in 'Animated film' category, 756 annotated in 'Operating system' category, 209 annotated in both categories, and 43 queries annotated only in 'Other' category.

5. Dataset Structure

In the delivered dataset, we produce a CSV file per task, containing the search sessions of all 100 participants (identified by anonymous codes), including the chronologically ordered submitted queries (consecutive duplicates have been removed), annotated according to the three binary categories defined above (see Table 3).

In addition, other files provide the following data:

- For each participant, their user id, age, gender, academic level (degree), assessment score for each of the four target domains, answer provided for each of the two tasks.
- For each search session (user id, task id), the complete history as shown in figure 2, including SERPs (e.g. row numbered 0 in the table) and visited URLs (e.g. row 6 of the table).
- For each SERP in every session, the HTML source of the web page. Some recorded SERPs are not in the history. It is possible that these SERPs stem from clicks on SERPs mentioned in the history or that they are noise produced by the recording tool.

Additional information includes: Ethics review committee certification, text of the agreement form signed by every participant, multiple-choice self-assessment questionnaire for the domain scores (per task, per participant) - see Section 3.4.1, prompts used for LLM annotation.

6. Analysis of Search Strategies

We present a first use case analysing variation in session structure through sub-objective search. Using query annotations, we categorise user behaviour based on the sequencing and distribution of sub-objectives within a session, as well as switching patterns between them. Although this analysis relies only on query data, the dataset enables substantially richer studies, including those incorporating visited page content. We consider different aspects of the search strategies.

First User's Query How do users begin their search session? Do they start their session by formulating a query related to both sub-objectives, or do they plan and sequence their search sub-objectives? In most cases, users begin sessions by submitting queries that refer to a single sub-objective. For TSF, only 5 users over 100 begin with queries belonging to both categories simultaneously, and 21 for TMR.

Participant	Id	Queries	Mythology	Art	Other
NV1162	1	renaissance peinture héraclès	1	1	0
NV1162	2	renaissance peinture héraclès femme	1	1	0
NV1162	3	renaissance peinture héraclès transformation	1	1	0
NV1162	4	héraclès	1	0	0

Table 3: Sample of automatic annotation for TMR

In addition, we examined whether the order of the sub-objectives in the instruction influences the first query. In other words, do users begin their session by formulating a query that referred to the first sub-objective mentioned in the instruction, or not? We measured a significant correlation (chi-square test, $p < 0.05$) for TSF, indicating that the participants who receive the task mentioning the "animated movie" sub-objective first formulate an initial query related to this sub-objective, and vice versa for operating systems. However, for TMR, the chi-square test proved to be non-significant ($p = 0.18$). However, it should be noted that no participant begins their session with a query related solely to the Renaissance if the instruction begins with Mythology.

Distribution of Sub-Objectives This sequencing may also appear throughout search sessions. In other words, is there a majority of queries referring to a single sub-objective, or two sub-objectives at the same time? Are all sub-objectives treated separately at least once in the session?

For TSF, an average of 14% of queries per session are "dual-topic" queries. 25 participants did not formulate any dual-topic queries. Sub-objectives are formulated individually in most cases. All but one user formulated at least one query annotated exclusively as belonging to the "Animated Film" category; the same pattern was observed for the "Operating System" category. For TMR, on average 45% of queries in a session are "dual-topic" queries, and all users formulate at least one to complete this task. 59 participants did not formulate a query solely related to the category "Art/Italian Renaissance"; only one individual who did not formulate any queries belonging solely to the category "Mythology".

This distinction in tasks may be explained by their structure. Indeed, for TMR, it is necessary to have the names of the protagonists in order to find a corresponding painting whose title is often composed of the names of the characters (e.g., Leda and the Swan).

Switching Between Sub-Objectives If the sub-objectives are formulated separately in the queries, do we observe many changes between the two sub-objectives? In other words, do users switch from one sub-objective to another? And if so, does this happen frequently?

We found an average of 2.4 changes per session for participants who searched for sub-objectives individually for TSF, and only 0.8 changes on average for TMR. Of these 98 participants for TSF, 6 of them never made any changes. In other words, users never formulated a query related to the 'Animated film' sub-objective and then directly to the 'Operating system' sub-objective, but they formulated a dual-topic query or an 'Other' query. Of the 40 participants who queried with a sub-objective only for the TMR task, only 23 made a change. These initial observations allow us to identify behavioural variations during web searches.

Visualisation of Search Strategies To explore further, we visualise search sessions based on the query annotations. These visualisations reveal differences in the users' strategies, i.e. behavioural distinctions at the session level. We have identified three search strategies inspired by the cognitive psychology literature: *logical*, *global*, and *exploratory* (Marchionini, 1995; Bates, 1990; Thatcher, 2006; Drabenstott, 2001), illustrated in Figures 4, 5, and 6.

In these figures, each dot represents a query whose colour and horizontal coordinate correspond to the automatic annotation of the sub-objective they belong to. Blue queries refer only to the first sub-objective, yellow queries only to the second, and green queries refer to both. Queries are organised vertically, with the first at the top.

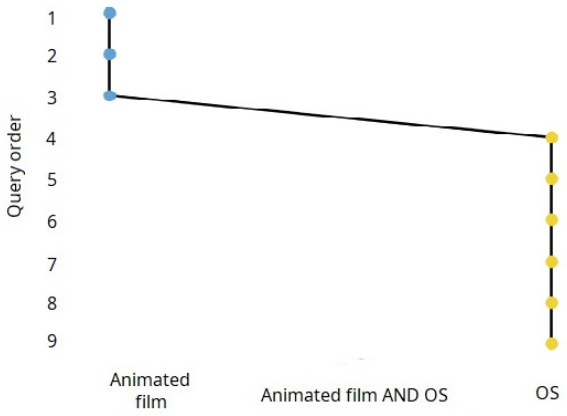

Figure 4: Representation of logical strategy

Figure 4 shows a strategy described as *logi-*

cal: the user sequences the sub-search objectives, starting by formulating three queries related to animated films, before moving on to operating systems by submitting six queries. Here, we observe a topical change between queries 3 and 4.

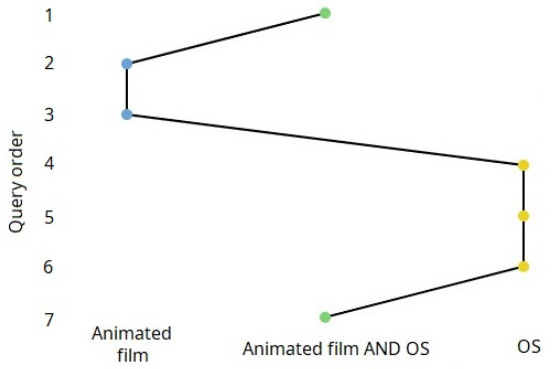

Figure 5: Representation of global strategy

In Figure 5, the user begins their search session directly with a dual-topic query (green dot) before performing the sequencing necessary to satisfy the information needs. The first query may indicate a lack of planning, distinguishing between logical and global strategies.

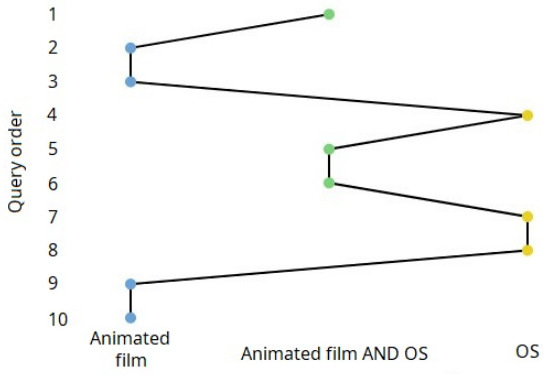

Figure 6: Representation of exploratory strategy

Session 3, in Figure 6, illustrates the *exploratory behaviour*. The user employs different query formulation tactics: queries related to a single sub-objective, dual-theme queries, at different stages of the session. In addition, we observe changes at different points in the session: at the beginning (see the change between queries 3 and 4) but also at the end of the session (see queries 8 and 9).

More detailed behaviour can be investigated, and more precise local descriptions are also made accessible by making use of the other relevant information (both local and global) in the dataset.

Limitations These initial analyses cannot be generalised to all web users. Indeed, the majority of the participants are young female students, it would be interesting to expand the panel of participants. This will enable us to observe if users with other socio-demographic characteristics use the same search strategies. It would be interesting to see whether socio-demographic characteristics generally influence search behaviour, as do the other elements collected (e.g. multiple-choice self-assessment questionnaire related to prior knowledge).

Another limitation of our dataset concerns the design of the search tasks. First, the tasks may lack realism, although the individual sub-objectives themselves reflect common information needs (e.g., searching for a movie by a particular film maker or a painting depicting a specific event). This reduced realism was a deliberate choice to enable a controlled analysis of the phenomenon under study. Furthermore, the tasks are confined to specific cultural domains. It therefore remains an open question whether similar variations would be observed with tasks addressing other types of topics.

7. Conclusion and Future Work

In conclusion, we develop a new French-language resource consisting search sessions involving 100 participants completing two multi-hop and multi-objective search tasks on a search engine. In total, we collected 3,162 queries that we fully annotated using LLMs according to formulated sub-search objectives. This enriched dataset opens ways for new types of research on how users handle complex search tasks, addressing a critical gap in existing resources by providing explicit goal decomposition traces alongside complete search histories.

Our first analyses were carried out using only queries. However, the PRIVaThe resource also provides the documents visited by users. These documents may refine the interpretation of the observed users' behaviours and help explain certain actions, such as sudden switches between sub-objectives. Another promising direction is linguistic variations, enabling a finer-grained characterisation of formulation tactics and a deeper understanding of search strategies. During data collection, we also recorded the verbalisations of 20 users. Although not yet available in the resource, they may offer valuable insight into users' intentions.

Finally, this dataset provides access to human reasoning when faced with multi-hop questions. It would be valuable to investigate whether certain forms of reasoning are better suited to respond to information needs. This approach would then enable comparison of human reasoning with that implemented by LLMs, potentially informing improvements to these models.

8. Appendix

Prompt used for the final TSF annotation

You are a trained linguist specialised in information retrieval, working as my assistant. Your task is to identify the semantic categories for search engine queries written in French by participants for an experience.

This is the research task given to participants: "Quel est le lien entre le long métrage d'animation réalisé par Pete Docter où Henry Guybet prête sa voix et la version 1.2 du système d'exploitation représenté par une spirale, associé à celui avec le pingouin ?"

And the expected response is: "La version 1.2 du système d'exploitation Debian représenté par une spirale, et associé à Linux, a comme nom "Rex". "Rex" faisant référence au personnage du film Toy Story doublé par Henry Guybet qui apparait également dans le film Monstres et Cie réalisé par Pete Docter."

For this annotation, you will consider only the following categories, based on the elements of the task and the answer:

A: Animated films including actors, directors, voices, films names, studios, characters

B: Operating system including versions, systems names

C: Other topics not covered

You will answer for each category with only "yes" or "no", indicating whether the topic is present or not in the query. You don't give any justification, only the category as in the example.

For example, if the query is "version 1.2 pinguoin spirale" the answers are:

A: no

B: yes

C: no

For example, if the query is "toy story version 1.2 debian wikipedia" the answers are:

A: yes

B: yes

C: yes

9. Bibliographical References

Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2850–2862, New York, NY, USA. Association for Computing Machinery.

Marcia J. Bates. 1990. Where should the person stop and the information search interface start? *Information Processing & Management*, 26(5):575–591.

Cheyenne Dosso, Jose G. Moreno, Aline Chevalier, and Lynda Tamine. 2021. Cost: An annotated data collection for complex search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4455–4464, New York, NY, USA. Association for Computing Machinery.

Karen M Drabenstott. 2001. Web search strategy development. *Online*, 25(4):18–27.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "you are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

Gemma Team and al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*, abs/2403.08295.

Eszter Hargittai. 2006. Hurdles to information seeking: Spelling and typographical mistakes during users' online behavior. *Journal of the Association for Information Systems*.

Laurie Hughes, Yogesh K Dwivedi, Tegwen Malik, Mazen Shawosh, Mousa Ahmed Albashrawi, Il Jeon, Vincent Dutot, Mandanna Appanderanda, Tom Crick, Rahul De', et al. 2025. Ai agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, pages 1–29.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*, abs/2310.06825.

Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, and Elena Simperl. 2019. Characterising dataset search—an analysis of search logs and data requests. *Journal of Web Semantics*, 55:37–55.

Llama Team and al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, abs/2407.21783.

Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering.

P. A. Nemkova, S. Ubani, and M. V. Albert. 2025. Comparing llm text annotation skills: A study on human rights violations in social media data. *arXiv preprint arXiv:2505.10260*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS marco: A human-generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268v3*.

Paul Owoicho, Jeff Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R Trippas, and Svitlana Vakulenko. 2022. Trec cast 2022: Going beyond user ask and system retrieve with initiative and response generation. In *TREC*.

Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Romina Sharifpour, Mingfang Wu, and Xiuzhen Zhang. 2023. Large-scale analysis of query logs to profile users for dataset search. *Journal of Documentation*, 79(1):66–85.

Joseph Sharit, Mario A. Hernández, Sara J. Czaja, and Peter Pirolli. 2008. Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the web. *ACM Trans. Comput.-Hum. Interact.*, 15(1).

Andrew Thatcher. 2006. Information-seeking behaviours and cognitive search strategies in different search tasks on the WWW. *International Journal of Industrial Ergonomics*, 36(12):1055–1068. Selected papers form CybErg 2005, the fourth International Cyberspace Conference on Ergonomics.

Xuanhui Wang and ChengXiang Zhai. 2007. Learn from web search logs to organize search results. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 87–94.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis. *International Journal of Corpus Linguistics*, 29(4):534–561.