Understanding and Modeling Heterogeneous Search Behavior

Nuha Abu Onq
nuha.abu.onq@student.rmit.edu.au
RMIT University
Melbourne, Victoria, Australia

Chenglong Ma
chenglong.ma@rmit.edu.au
RMIT University
Melbourne, Victoria, Australia

Mark Sanderson
mark.sanderson@rmit.edu.au
RMIT University
Melbourne, Victoria, Australia

Falk Scholer
falk.scholer@rmit.edu.au
RMIT University
Melbourne, Victoria, Australia

Abstract

We investigate *between-user* drivers of query variability through a controlled between-subject, full-factorial user study that manipulates *age*, *gender*, and *language proficiency* across six backstory-driven search tasks. From initial queries, session logs, and post-task interviews, we quantify how demographic and task factors shape query-, task-, and session-level behaviors. We further derive a small set of interpretable latent search dimensions from user evidence to analyze and simulate heterogeneous query behavior. Our results show that age is the most consistent predictor of query formulation and search interaction patterns. Gender and language differences are more selective, and task context is further associated with these patterns. The latent dimensions help explain the differences as variation in search strategy rather than uniform differences in engagement or ability. The paper provides a trait-informed view of heterogeneous search behavior that supports more user-aware analysis and robustness-oriented evaluation in IR.[1]

CCS Concepts

• **Information systems** → **Information retrieval**; **Information retrieval query processing**; **Query representation**.

Keywords

Query formulation, Search behavior, Latent user modeling

ACM Reference Format:
Nuha Abu Onq, Chenglong Ma, Mark Sanderson, and Falk Scholer. 2026. Understanding and Modeling Heterogeneous Search Behavior. 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, 13 pages. https://doi.org/10.1145/3805712.3809618

1 Introduction

People exhibit substantial diversity in how they formulate search queries, when seeking the same information need [4]. Such variability reflects differences in underlying search strategies [2] and can influence how retrieval systems interpret intent [4, 13, 24], what is retrieved [5, 9, 13], and how robust effectiveness is under heterogeneous query strategies [58]. Understanding the origins of diversity, and how it relates to user characteristics, remains a central challenge for user-aware and fair Information Retrieval (IR).

Past Query Variant (QV) studies [8, 51] highlighted the diversity. However, the method of collection (crowd workers) abstracted the exploratory nature of real–world search and lacked demographic or user–level interaction data. Digital library logs record interaction patterns [36] rather than the linguistic [1] or behavioral diversity of queries [2, 8, 51]. Recently, Large Language Models (LLMs) have been used to generate diverse queries or simulate users, but many approaches rely on manually specified [83] or demographically sketched personas [3]. Neither derive user characteristics from empirical evidence, limiting interpretability and validity.

Such gaps motivate the question: *how do observable user characteristics and latent search mechanisms shape initial query formulation and early search behavior?*

We address the question through a controlled user study with an evidence-based simulation framework. A $2 \times 2 \times 2$ factorial study with 64 participants manipulates age, gender, and language proficiency across six search tasks spanning three levels of cognitive complexity [2, 80]. We collect interaction logs and post-task interviews, focusing on between-subject variability of queries. We introduce an LLM-based profile–simulation pipeline that infers latent search traits from user evidence and generates task-specific queries under profile constraints. We further define a set of IR-specific latent mechanisms that are (i) separable, (ii) inferable from logs and interviews, and (iii) can be employed in simulation.

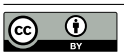
Our design enables two analysis levels: **observable**, quantifying how demographics and task characteristics shape linguistic properties of initial queries and early interaction patterns; and **latent**, testing whether evidence-grounded trait profiles provide an explanation for stable between-user differences, and whether persona conditioning improves simulation fidelity beyond simple baselines. Together, these analyses form a structured research pipeline that progresses from empirical characterization to latent inference and finally to explanatory validation.

Our study is guided by the following research questions:

RQ1. How do demographic characteristics influence query formulation and search behavior?

- **H1** **Age**: Older users use more newly added keywords, generate greater query variety, formulate longer, more complex queries, expect more queries and documents, access more documents, and spend longer completing tasks.

[1]The code is available at https://github.com/ChenglongMa/qv-lab.



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809618

H2 **Gender**: Females show selective differences in interaction preferences (e.g., Search Engine Result Page (SERP) click behavior) on the same dependent measures.

H3 **Language Proficiency**: Multilingual users show differences in semantic alignment and exploration behavior on the same dependent measures.

This RQ captures observable, demographic-level variation in initial query formulation and early interaction behavior.

RQ2. How to infer stable, interpretable, mechanism-oriented latent search traits from empirical evidence?
This RQ focuses on methodological modeling, where observed queries, interaction logs, and post-task rationales are used to infer latent search mechanisms and user profiles.

RQ3. What mechanism-oriented latent search dimensions emerge, and how do they influence user observable search behavior?
This RQ evaluates the explanatory and predictive validity of the inferred traits, including whether they account for stable between-user differences and improve behavior modeling compared to non-personalized baselines.

This paper contributes: 1) A hybrid, evidence-gated pipeline for categorizing novel terms in QVs into a taxonomy, supporting strategic analysis at scale. 2) A profile-and-simulation framework that infers latent search traits with traceable evidence, and uses them for controlled LLM-based query generation. 3) An integrated analysis linking observable demographics, task effects, and inferred latent mechanisms to explain heterogeneous query formulation, with implications for user-aware and robustness-oriented IR evaluation.

2 Related Work

This section reviews prior work on factors influencing query formulation, user-aware query variants, and LLM-based user simulation.

2.1 Factors Influencing Query Formulation

Demographic characteristics (e.g., age, gender, language proficiency) shape query formulation, motivating research on demographic signals in queries and their implications for user-aware or fairness-aware IR [79, 81]. **Age**: older users tend to employ less flexible strategies and emphasize result evaluation over iterative query refinement [14, 19, 26, 27, 69], younger users adopt more diverse strategies for certain needs [77]. **Gender** appears to effect query frequency/length and engagement patterns [38, 41, 45, 47, 52, 56, 62, 69, 86], though age can be a stronger driver than gender alone [69]. **Language proficiency** affects effort allocation and reformulation: non-native speakers often spend more time and reformulate more, and multilingual users may switch languages depending on proficiency and task context [16, 21, 22, 44, 48, 71, 72, 85]. Much of the literature examines factors in isolation or outside controlled shared information needs, leaving *between-user variability* in initial formulations comparatively under-characterized.

2.2 User-aware QVs

QVs are widely used to probe system robustness under heterogeneous expressions of the same underlying need, often via collections of crowd-sourced QVs [4, 8, 9, 13, 24, 51, 58]. Beyond crowd collection, variants can be generated automatically through behavioral signals such as click graphs [46] or via neural/LLM-based generation [5, 17]. Recent work further explores *user-aware* conditioning: demographically inspired prompts can substantially shift rankings and apparent robustness [3], while comparisons between LLM-generated and human (or crowd-worker) QVs reveal persistent gaps in how well current LLMs capture human query expression patterns [6, 83]. However, QVs are often treated primarily as evaluation artifacts or synthetic inputs rather than as manifestations of real users' strategies under controlled needs, and the demographic determinants of *initial* QV diversity remain underexplored.

2.3 User Simulation with LLMs

User modeling and simulation have long supported evaluation without explicit relevance judgments, from click-based preference inference [37] to rule-based session simulators that model interaction loops and their impact on evaluation [11]; a survey summarizes this landscape and its role in evaluation [10]. LLMs have been adopted to simulate session-level behaviors. USimAgent [84] uses an LLM to jointly simulate querying, clicking, and stopping. Work in conversational search introduces parameterized simulators and multi-trait decoding schemes [67]. A limitation is that the traits are typically manually specified, and the link between simulated behavior and real user characteristics is indirect. This motivates simulation frameworks that ground trait construction and conditioning in empirical evidence, enabling more interpretable and testable connections between user differences and generated query behavior.

3 Methodology

This section describes the user study design, data collection, and modeling framework for simulating query formulation behavior.

3.1 User Study

We used a $2 \times 2 \times 2$ between-subject full-factorial study to examine how demographic characteristics influence search behavior.[2] The manipulated variables were:

- **Age**: Young (18–42) vs. Old (50+)
- **Gender**: Male vs. Female
- **Language proficiency**: Monolingual (native English) vs. Multilingual (non-native)

The design produced eight unique demographic groups (Old—Male—Monolingual, Young—Female–Multilingual, etc.), with eight participants per group. The design ensured balanced representation across all factor levels (32 participants per level for each variable) and enabled reliable estimation of main and interaction effects.

3.1.1 ***Setting.*** All study sessions were conducted in a controlled laboratory environment to minimize external variation. Sessions took place in the same room with consistent lighting, noise levels, temperature, and absence of visual distractions. Participants used identical workstations, signed-out browser configurations, and network connections, and received standardized task instructions. No adaptive system features were introduced. Each session lasted approximately 45–60 minutes, and participants received a $65 (AUD) incentive for their participation. Participants were recruited through multiple channels, including flyer distribution, a

[2] Approved by the RMIT Human Research Ethics Committee (Project Id: 27520).

recruitment organization, and researchers' personal and professional networks. The approaches enabled outreach across a broad population, covering different occupations and age groups. The company supported the recruitment of older participants, who are typically underrepresented in such studies.

3.1.2 ***Participants.*** A total of 64 participants completed the study. Eligibility was determined using a screener questionnaire, which also assigned participants to one of the eight demographic groups. Each group included eight participants, ensuring balanced coverage across age, gender, and language proficiency.

We focus on widely studied demographic variables for which reliable operationalizations exist in controlled experimental settings. Given the full-factorial design and associated recruitment challenges, restricting participants to the two most represented groups enabled adequate sample sizes per condition while maintaining comparability with past work. **Age** groups were defined based on ranges reported in prior studies on age-related differences in search behavior [15, 20, 65, 66, 68, 73], using their lower and upper bounds to define *younger* (18–42) and *older* (50+) cohorts. Our sample size enabled the use of broader, literature-informed groupings while maintaining sufficient statistical power. **Gender** was operationalized as a binary factor, which has been adopted in the past [31, 76, 81]. For **Language Proficiency**, *monolingual* participants were defined as those speaking English as their only language, while *multilingual* participants reported speaking more than one language and being proficient in English. Those not proficient in English were excluded from the study, as the experimental tasks and materials were conducted entirely in English.

3.1.3 ***Procedures.*** The following fixed sequence of questionnaires were administered using Qualtrics.

(1) **Screener:** Used in recruitment, eligibility verification, and quota management of demographic groups.
(2) **Demographics:** At the beginning of the session we collect participant demographics and search experience.
(3) **Pre-task:** Participants viewed a task backstory (as an image to prevent copying) and reported expected number of queries and documents, interest, and familiarity. Then, they formulated their initial search query directly in Qualtrics, avoiding exposure to search engine autocomplete suggestions.
Participants completed the same six information-seeking tasks selected from the UQV100 [8]. The tasks order was randomized to mitigate learning and order effects. The tasks covered varying cognitive complexity, including two *Remember* (easy), two *Understand* (moderate), and two *Analyze* (complex) tasks, following established cognitive complexity classifications [2] (task IDs: 203, 206, 224, 225, 293, 296).
(4) **Search Session:** Using a clean browser state, participants submitted their initial query via a provided Google search link and interacted with the SERP, including clicking results, navigating to result pages, reformulating queries if needed, and examining documents until satisfied.
(5) **Post-task:** Participants reported post-task satisfaction, perceived difficulty, and familiarity with the retrieved information.
(6) **Exit:** Upon completing all tasks, participants reflected on overall task complexity and search experience.

Post-task Interview. A semi-structured interview elicited deeper insights into participants' search strategies, decision-making processes, and query formulation behavior. Recorded using Microsoft Teams, the interview addressed the following aspects:

- **Rationale for novel terms.** Participants were asked to explain the inclusion of specific terms in their queries, distinguishing between terms that replaced existing words and those that were newly introduced.
- **Query length considerations.** They were asked to discuss the factors influencing their query length.
- **Query formulation strategy.** They were asked to describe their decision to formulate queries as either keyword– or question–based query, and to reflect on the extent to which their prior knowledge of the task influenced term selection.
- **Comparison of expected and actual search behavior.** They were asked to evaluate whether their initial expectations regarding the number of queries and documents required aligned with their actual search experience.
- **Complexity assessment rationale.** They were asked to provide reasoning behind the complexity ratings they assigned to the search tasks.

3.1.4 ***Data Collection and Analysis.*** Collected data included questionnaire responses, interview transcripts, observational notes, and logged search interactions (e.g., queries, clicks, and navigation behavior). Across 64 participants, each completing six study tasks described in the *Pre-task questionnaire*, we obtained 384 QVs. Self-reported data from pre- and post-task questionnaires were collected using Likert scales, numerically encoded for analysis (1–5, with higher values representing greater familiarity, interest, satisfaction, or perceived difficulty). Quantitative analyses examined the main effects of age, gender, and language proficiency using a full-factorial design. For two-level factors, independent-samples t-tests were used; task complexity (more than two levels) was analyzed with one-way ANOVA followed by Tukey's HSD post hoc tests. Chi-square tests of independence assessed differences across participant groups in QV counts or categories, query form (question vs. topic), sentiment, and part-of-speech (POS) patterns. Qualitative analyses of interviews contextualized behavioral patterns and helped explain demographic differences in task approaches, complementing quantitative measures. Table 1 summarizes the dependent variables analyzed at the query, session, and task levels. For linguistic feature extraction, POS tagging was performed using spaCy's `en_core_web_trf` model (v3.8). Sentiment was computed using the VADER sentiment analyzer. Query type was determined heuristically: queries were classified as *questions* if they began with common interrogative keywords (e.g., *what*, *how*, *why*) or ended with question marks; all remaining queries were classified as *topic*.[3]

3.2 Hybrid Novel Term Categorization

To support strategy-level analysis of QVs, we map each novel term to a structured taxonomy that distinguishes whether the term modifies the task description or enriches it with additional constraints.

[3]Data will be made available upon request.

Table 1: Dependent variables organized at the query, task, and session levels.

Level	Metric	Description
Query	QVs	Total unique queries, average query length.
	Novel Terms	Number of new terms, associated categories.
	Linguistic Features	Jaccard similarity, Flesch Reading Ease (FRE), semantic similarity, lexical density, sentiment, query type, POS patterns.
Session	Dwell Time	i) **SERP dwell time:** Time on the initial SERP evaluating the results before clicking. ii) **Result dwell time:** Time spent on a clicked result page, measured from the moment the user navigates from the SERP to that page.
	Clicks	i) **SERP:** Count of clicks on SERP results. ii) **Follow-up:** Clicks after the initial SERP click, e.g., navigation to additional pages. iii) **Question:** Clicks on Google's "People Also Ask" questions. iv) **Image:** Clicks on image results. v) **Product:** Clicks on product/shopping listings.
	Tab Navigation	Number of tab switches (e.g., Forums, Images, Shopping).
Task	Familiarity	Prior familiarity with a task.
	Interest	Participant interest in a task.
	Difficulty	Self-reported difficulty of a task.
	Satisfaction	Satisfaction with retrieved information.

This categorization serves two purposes in our study: it enables large-scale analysis of how users translate shared information needs into a query, and it provides a category-level representation used later in simulation evaluation. To operationalize the QV taxonomy of Abu Onq et al. [2] at scale, we develop a hybrid pipeline that combines deterministic evidence extraction with constrained LLM inference. Given a task backstory x_t, a user query u_q, and a detected novel term u_t that appears in u_q but not in x_t, the pipeline assigns a two-level label. At the first level, *Modification* indicates that u_t plausibly replaces a specific lexical unit in x_t while preserving the underlying information need, whereas *Enrichment* indicates that u_t adds an extra constraint or expansion beyond the backstory. At the second level, we refine each first-level decision into a more specific subtype. Under *Modification*, the subtype space includes *Operator* (e.g., "+"), *Misspelling*, *Abbreviation*, *Transformation*, *Antonym*, and *Semantic replacement*. Under *Enrichment*, the subtype space includes *Imperative*, *Resource*, *Information-type* (e.g., "tutorial"), *Entity*, *Modifier*, and *Off-topic*.

Evidence Extraction and Gating. Because many novel terms admit multiple plausible interpretations, we first extract candidate replacement evidence before invoking the LLM. We construct a backstory lexicon $B(x_t)$ consisting of short content phrases and words. For each candidate span $b \in B(x_t)$, we compute string- and semantics-based cues: (i) lemma/stem match, (ii) normalized edit distance, (iii) character n-gram overlap, (iv) embedding cosine similarity, and (v) lexical relations. We organize these cues into reliability tiers so that high-precision signals can guide the decision more strongly than heuristic ones. *Strong* signals are high-precision indicators (e.g., exact lemma/stem match, edit distance ≤ 1 for content words, or unambiguous operator patterns); *Medium* signals are informative but potentially ambiguous (e.g., approximate string similarity above a threshold); and *Weak* signals are heuristic (e.g., embedding similarity or broad lexical relations), used only as supporting context.

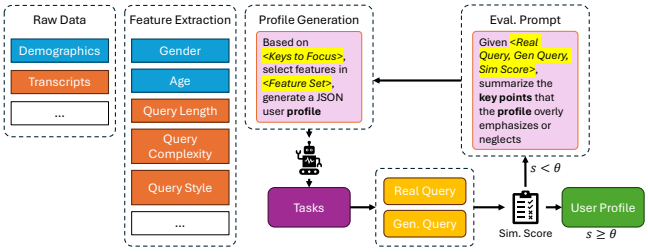


Figure 1: Profile-simulation-evaluation framework overview.

To reduce noise propagation, we gate the evidence before passing it to the LLM. Concretely, we retain only candidates that satisfy fixed thresholds and a score-margin criterion:

$$C(u_t) = \left\{ b \in B(x_t) \;\middle|\; s_{\text{str}}(u_t, b) \geq \tau_{\text{str}} \wedge s_{\text{str}}(u_t, b) - s_{\text{str}}(u_t, b^{(2)}) \geq \delta \right\},$$

where s_{str} is a composite string score derived from (i)–(iii), $b^{(2)}$ is the second-best match under s_{str}, and τ_{str} and δ are fixed on a small development subset and then held constant across all experiments. We output at most top-k candidates (e.g., $k = 3$). If $C(u_t) = \emptyset$ and no *Strong* replacement signal is present, we emit `no_reliable_backstory_match` to explicitly indicate insufficient replacement evidence.

Constrained LLM Decision. The LLM then makes the final taxonomy decision using the task backstory, the query, the novel term, and the gated evidence. The model is given a fixed label space and a decision constraint: taxonomy definitions from Abu Onq et al. [2] are treated as authoritative, while evidence cues are treated as heuristics that may be rejected if inconsistent with context. This design lets the model use context to resolve ambiguous cases, while limiting arbitrary label assignments. We implement a replacement test at the first level: if any *Strong* replacement signal exists, the LLM assigns *Modification*; otherwise it assigns *Enrichment*. Conditional on the first-level category, the LLM selects a second-level subtype using the provided evidence and context, preferring higher-precision subtypes when applicable. The output is a structured record with `first_level`, `second_level`, `confidence`, and `used_evidence`.

3.3 LLM-based User Search Profile Simulation

We use a profile–simulation framework to test whether behaviorally grounded latent traits help explain and reproduce between-user differences in query formulation. The framework takes demographic attributes, interview transcripts, and interaction logs as evidence, infers a structured user profile, and then conditions query generation on that profile. We treat transcript-enhanced profiling as a high-information analysis setting rather than a required deployment assumption; accordingly, we also evaluate reduced-evidence conditions through controlled ablations. Figure 1 summarizes the full workflow of profile generation and evaluation.

Inputs and Evidence Representation. For each user u, we observe: (i) demographic attributes D_u (e.g., age group, gender, language proficiency), (ii) an interview transcript T_u collected after all tasks, and (iii) a task set $X_u = \{x_t\}_{t=1}^{6}$ together with the corresponding human queries and interaction logs L_u. These sources play complementary roles: demographics provide coarse observable group information, transcripts provide post-hoc rationales, and logs provide observed behavioral evidence. Our profile generator consumes a configurable evidence pack determined by switches over sources, $E_u^{(s)} \subseteq \{D_u, L_u, T_u\}$, which lets us test how much each source contributes to profile quality and simulation fidelity.

Because raw transcripts contain procedural content and redundancy, we compress T_u into trait-aligned evidence snippets before profile inference. We first chunk T_u into paragraphs, filter common non-evidence segments (e.g., operational instructions), and retrieve top-K snippets per latent trait using a configurable retriever (BM25 by default; optional embedding retrieval). Each snippet is stored with provenance (document name and paragraph index), yielding a transcript evidence pack $T'_u = \{T^K_{u,t}\}_{t=1}^{6}$. This step reduces omission errors while preserving traceable evidence for later auditing.

From the interaction logs, we construct two complementary representations so that the profile can reflect both within-task behavior and more stable user tendencies. First, task-level sequences $L^{\text{seq}}_{u,t}$ store the submitted queries for task t in timestamp order, together with behavioral fields such as suggestion/autocomplete use, click counts, dwell times, and tab/page navigation. Second, user-level summaries L^{sum}_u capture robust aggregates (e.g., median/IQR and rates) and a small set of derived proxies aligned to the latent traits, such as suggestion usage rates, dwell-time shares, click-type diversity, and sequence-coherence indicators. The sequential view, i.e., $L_u = \{L^{\text{seq}}_{u,t}\}_{t=1}^{6} \cup L^{\text{sum}}_u$, preserves temporal structure, whereas the summary view exposes cross-task regularities.

Given an evidence condition c, we provide the profile generator with a bounded-length evidence pack

$$E_u^{(c)} = (\mathbb{I}[\text{Demo} \in c] D_u,\ \mathbb{I}[\text{Logs} \in c] L_u,\ \mathbb{I}[\text{Trans} \in c] T'_u,\ X_u^{\text{train}}),$$

where X_u^{train} denotes the subset of tasks used for profile inference under a holdout protocol.

Profile Inference and Trait Construction. We constructed this trait space as an intermediate behavioral layer between coarse demographic attributes and observable QVs. Prior work has shown that demographic conditioning alone, or manually specified personas, provides only an indirect account of user-specific query variation [3, 83]. We therefore define traits that capture search mechanisms that can plausibly produce different query formulations under the same information need. Concretely, each user is associated with a normalized trait intensity score vector

$$z_u = (z_{u,1}, \ldots, z_{u,6}) \in [0,1]^6,$$

corresponding to Evidence Rigor (ER), Effort Willingness (EW), Exploratory Curiosity (EC), System Reliance (SR), Strategy Formation (SF), and Authority Preference (AP). The dimensions reflect recurring distinctions in information seeking and search behavior, including effort allocation, exploratory breadth, planning versus reactive interaction, reliance on system affordances, and credibility-oriented source evaluation. They are further motivated by the observation that similar surface behavior may arise from different underlying drivers [49, 54], and by depth–breadth trade-offs in information processing [43]. While personality theory provides an interpretive lens for stable individual differences [50, 63], we use these dimensions as analytic constructs for modeling observed search behavior, not as psychometric measures of stable personality. We retained only dimensions that are conceptually motivated, inferable from logs or post-task rationales, and useful for interpretable grouping and controlled simulation.

The profile generator consumes $E_u^{(c)}$ and outputs a structured profile $P_u^{(c)}$ containing: (i) user-level trait scores z_u, (ii) per-trait confidence scores, and (iii) evidence pointers that reference concrete log-derived indicators and transcript snippets. To reduce over-inference, we impose a minimum evidence requirement per trait: when transcript evidence is enabled, each trait must be supported by at least one log-based item and one transcript snippet; otherwise, multiple log-based items are required. Missing transcript coverage is treated as reduced confidence rather than contradictory evidence.

We further stabilize the inferred profile through repeated sampling under the same evidence pack. Specifically, we sample n independent profiles and aggregate them into a consensus profile $\bar{P}_u^{(c)}$ by averaging scalar trait values and majority voting categorical auxiliaries; unstable fields are flagged as low-confidence. We report 95% uncertainty intervals derived from cross-sample variability rather than relying on intervals produced directly by the LLM. The six traits are defined as follows:

- **Evidence Rigor**: Degree to which a user demands and verifies evidence before acceptance.
- **Effort Willingness**: Investing time/effort to improve outcome quality (deep reading, query crafting), beyond task demands.
- **Exploratory Curiosity**: Preference for broader information gain, exploring alternatives, perspectives, or related knowledge beyond the minimum answer.
- **System Reliance**: Delegation of querying and navigation decisions to system mechanisms (suggestions, autocomplete, SERP modules), rather than self-directing.
- **Strategy Formation**: Extent of plan-first behavior, i.e., forming an initial strategy (keywords/sub-questions/path) before interaction, versus reactive trial-and-error.
- **Authority Preference**: Default preference to treat institutional/expert sources as credible warrants (authority cues as a legitimacy prior), distinct from verification rigor.

To reduce systematic misclassification, these traits are interpreted with explicit guardrails. For example, high activity alone is not sufficient evidence of *Evidence Rigor*; *Authority Preference* must not inflate *Evidence Rigor* without verification evidence; and suggestion/autocomplete usage primarily supports *System Reliance*.

Persona-conditioned Simulation. Given a consensus profile $\bar{P}_u^{(c)}$ and a task backstory x_t, the simulator A_{sim} generates a sequence of synthetic queries

$$\hat{Q}_{u,t} = \{\hat{Q}^{(1)}_{u,t}, \ldots, \hat{Q}^{(n_{u,t})}_{u,t}\},$$

where $n_{u,t}$ matches the number of submitted human queries observed for user u on task x_t under position-wise alignment. The simulator is instructed to follow the inferred trait intensities, knowledge bounds, and query style constraints while producing realistic

web-search queries. We keep generation single-shot and disallow external retrieval or document access so that any improvement can be attributed more directly to the inferred profile rather than to additional information sources.

Evaluation and Refinement. We evaluate the generated queries against held-out human behavior, using the simulation step as a controlled test of whether the inferred profile captures behaviorally meaningful regularities. For each task x_t, we compute embedding-based cosine similarity between the human query sequence $Q_{u,t}$ and the generated sequence $\hat{Q}_{u,t}$ under position-wise alignment, and report both task-level and user-level aggregates. To reduce overfitting, profiles are inferred from a subset of tasks X_u^{train} and evaluated on disjoint holdout tasks X_u^{hold}. We also record the lowest-similarity mismatches per task as diagnostic exemplars for refinement.

The evaluation agent A_{eval} then reads $(x_t, Q_{u,t}, \hat{Q}_{u,t}, \bar{P}_u^{(c)})$ together with the similarity metrics and an audit of the evidence pack, and produces a structured critique. The critique identifies (i) which traits appear over- or under-emphasized relative to the observed behavior, (ii) which evidence may have been missed or misweighted, and (iii) actionable patches to evidence selection and prompting. If refinement occurs, the profile generator updates $\bar{P}_u^{(c)}$ under two safeguards: (a) consistency filtering, which samples multiple critiques and applies only stable, metric-supported suggestions, and (b) anti-leakage constraints that prevent human queries or transcript segments from being injected verbatim as "evidence" into the profile. These safeguards are intended to reduce closed-loop bias and to keep refinement tied to auditable evidence rather than unconstrained model self-correction.

Refinement iterates up to a maximum number of rounds and terminates early if (i) holdout similarity exceeds a predefined threshold θ, (ii) improvement falls below ϵ for consecutive rounds, or (iii) A_{eval} judges the profile self-consistent and sufficiently supported by evidence. The final output of the framework is the refined consensus profile, which we use for downstream analysis and controlled query simulation.

4 Results

This section reports observed differences in query formulation and search behavior across demographic, task, and inferred trait factors.

4.1 User Experiment Analysis

We examined the influence of demographics and task characteristics on both query formulation and search behaviors (see Table 2).

Effects of Demographics. Demographics influence various aspects of queries and search interactions, yet participants from all groups generated many queries across tasks.

Age influenced multiple aspects of query formulation: older participants produced longer queries, both raw and processed, with more punctuation and novel terms, and expected documents. In contrast, younger participants' queries were semantically denser with higher lexical overlap, suggesting greater alignment in conceptual expression. Age also affected search behavior: older users spent more time examining SERPs, whereas younger participants engaged more actively with search results, clicking more frequently on SERPs, images, and navigating more tabs.

Gender effects were more subtle. Males formulated longer queries with higher punctuation counts and included more novel terms, but there were no significant differences in semantic similarity or lexical density. Females, on the other hand, spent more time on SERPs, clicked more on images and question links, and demonstrated slightly higher engagement with tab navigation.

Language proficiency was associated with query-level differences: monolingual participants produced queries with higher semantic similarity and greater use of punctuation and novel terms. In contrast, multilingual participants tended to issue more expected queries and documents and engage more with images and tab navigation, suggesting a broader exploration of the search process.

No significant effects of demographic factors were observed for query readability, and sentiment, result-page dwell time, or product clicks, indicating that these behaviors were consistent across groups. This suggests that observed variability may reflect individual-level differences rather than systematic group-level effects.

Effects of Task Characteristics. Task-level factors had strong and consistent effects on both query formulation and search interactions. Task complexity influenced query diversity, length, and the introduction of novel terms: more cognitively demanding tasks (*Analyze* > *Understand* > *Remember*) led to longer, more complex queries, whereas simpler tasks *"Remember"* elicited higher semantic similarity, indicating greater alignment in how users interpreted the task. Similarly, task length affected query coherence: queries for shorter tasks were more semantically and lexically similar, suggesting that concise tasks promote consistent and simple query formulation, whereas longer tasks elicited greater diversity, length, and inclusion of novel terms. Session behaviors mirrored these patterns. Complex and long tasks were associated with longer SERP dwell times, more question clicks, increased SERP interactions, and extensive follow-up activity, reflecting higher search effort. In contrast, shorter and simpler tasks produced more uniform queries and interactions, with greater alignment in image and product clicks.

Overall, these findings indicate that task and demographic characteristics influence both query formulation and search behavior.

In addition to behavioral measures, we examined subjective task-level outcomes—familiarity, interest, satisfaction, and perceived task complexity—across tasks that do not require specialized domain knowledge. While Table 3 presents the quantitative patterns, participants elaborated on these experiences in interviews, offering insights into their reasoning and motivations behind their search behaviors. Across tasks, age, gender, and language proficiency showed consistent differences with these subjective measures.

Older participants reported higher familiarity, interest, and satisfaction across several tasks. For example, in T_{224} (soup recipe), prior cooking experience increased their interest and satisfaction, whereas in T_{206} (wind power), broader life experience supported higher prior familiarity with fact-based complex topics. In contrast, younger participants rated T_{293} (Facebook evidence) and T_{296} (battery recycling) as more complex, noting that T_{296} required unfamiliar procedural problem-solving, while T_{293} demanded critical thinking to evaluate social claims and online evidence, which prompted them to generate more diverse queries than older groups. These complexity ratings reflect differences in prior knowledge and real-world experience, highlighting how age shapes the perception of task difficulty and engagement.

Table 2: Statistically significant effects of user and task factors on dependent variables across query, task processing, and session levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). *Task Complexity* denotes complexity level (*A* = Analyze, *U* = Understand, *R* = Remember). For *Query Form*, *Que./Top.* = Question/Topic. For *Sentiment*, *Neu./Pos./Neg.* = Neutral/Positive/Negative.

Level	Dependent Variable	Age	Gender	Language	Task Complexity	Task Length
Query	QVs (Count)	—	—	—	A > U **, A > R ***, U > R ***	Long ***
	Query Length (Word)	Old **	Male **	—	A > U ***, A > R ***	Long ***
	Processed Query Length	Old **	Male ***	—	—	Long ***
	Punctuation Count	Old ***	Male **	Monolingual *	—	—
	Novel Terms (Count)	Old **	Male **	Monolingual *	A > R ***, U > R ***	Long ***
	Search Strategy (First-level)	Young ***	Male **	—	A > U **, A > R **	Long ***
	Search Strategy (Second-level)	Young ***	Male ***	Multilingual ***	A > U **, A > R **	Long ***
	Jaccard Similarity	Young *	—	—	R > A ***, R > U ***	Short ***
	Semantic Similarity	Young **	—	Monolingual *	R > A ***, R > U ***	Short ***
	POS Pattern	—	—	—	A > U ***, A > R ***, U > R ***	Long ***
	Expected Queries	—	—	Multilingual *	A > U ***, A > R ***	Long ***
	Expected Documents	Old *	—	Multilingual *	A > U ***, A > R ***	Long ***
	Lexical Density	Young ***	—	—	R > A ***, U > A ***	Short **
	FRE	—	—	—	U > A ***, R > A ***, R > U **	Short ***
	Query Form	—	—	Que. Multilingual *	Que. A > R ***, U > R ***	Que. Long ***
				Top. Monolingual *	Top. R > A ***, R > U ***	Top. Short ***
	Query Sentiment	—	—	—	Neu. U > A ***, R > A *** Pos. A > U ***, A > R *** Neg. A > U **, A > R ***, U > R *	Neu. Short *** Pos. Long *** Neg. Long **
Task Proc.	First Click	—	—	Multilingual **	A > U ***, A > R ***, U > R *	Long ***
	Last Click	—	—	Multilingual **	A > U ***, A > R ***	Long ***
	Page Submit	—	Female *	—	A > U ***, A > R ***, U > R *	Long ***
	Click Count	—	—	—	A > R **	—
Session	SERP Dwell Time	Old ***	Female ***	Multilingual *	A > R ***, U > R **	Long **
	Result Pages Dwell Time	—	—	—	A > R ***, U > R ***	Long **
	SERP Clicks	Young **	—	—	A > R ***, U > R **	Long **
	Follow-up Clicks	—	—	—	U > A ***, U > R ***	—
	Image Clicks	Young *	Female *	Multilingual *	R > A ***, R > U ***	Short ***
	Question Clicks	—	Female *	—	A > U **, A > R ***	Long ***
	Product Clicks	—	—	—	R > A **, R > U **	Short **
	Tabs Navigated	Young ***	—	Multilingual ***	R > A ***, R > U ***	Short ***

Females reported greater interest and familiarity with everyday or experiential tasks, such as T_{224} (soup recipe). Although gender did not significantly affect interest or familiarity for T_{225}, males generated more diverse queries, suggesting that query diversity reflects cognitive effort. Males also showed higher interest and post-task familiarity for T_{206}, likely reflecting knowledge or confidence in fact-based evaluations. Multilingual participants reported higher familiarity and interest in tasks requiring broader knowledge integration (e.g., movie reviews, online shopping, and battery recycling), benefit from their diverse linguistic and cultural experience.

Across all tasks, prior familiarity was positively associated with satisfaction (Spearman's $\rho = 0.23$, $p < 0.001$), indicating that users with greater domain knowledge report more positive task experiences. Interest showed a similar positive relationship with satisfaction ($\rho = 0.20$, $p < 0.001$), whereas perceived task complexity was strongly negatively associated with satisfaction (ρ = -0.53, $p < 0.001$), highlighting the joint role of cognitive and affective factors in shaping perceived task success. Despite these differences in subjective experience, we observed no significant differences between familiar and unfamiliar users in query diversity, measured by the number of original and processed queries. This suggests that familiarity shapes perceptions but not the diversity of queries.

4.2 User Simulation Validation

We report a lightweight simulation sanity check to ensure that persona conditioning does not produce degenerate query outputs. Results focus on the inferred profiles, their stability, and the user groupings induced by profile features.

Novel Term Categorization Reliability. To validate the categorization reliability, two trained annotators evaluated agreement on a held-out validation set of 100 novel-term instances sampled to reflect the task distribution. Inter-annotation agreement was

Table 3: Significant differences in task-level subjective ratings across participant groups and tasks (T_i denotes Task i).

Task Topic	Dependent Variable	Dominant Group
#203 Les Misérables Reviews	Familiarity(Before)	Multilingual**
	Interest	Multilingual**
#206 Wind Power	Familiarity(Before)	Old**
	Interest	Male*, Multilingual*
	Familiarity(After)	Male**
#224 Chicken Soup Recipe	Familiarity(Before)	Female*
	Interest	Old**, Female**, Multilingual*
	Familiarity(After)	Female**
	Satisfaction	Old*, Female**
#225 Black and Gold Shoes	Familiarity(Before)	Multilingual**
	Interest	Multilingual***
	Task Complexity	Old*
#293 Facebook Evidence	Task Complexity	Young*
#296 Car Battery Recycling	Familiarity(Before)	Old***, Multilingual*
	Interest	Old**, Multilingual*
	Familiarity(After)	Old***
	Satisfaction	Old**
	Task Complexity	Young**

very high, with Cohen's $\kappa = 0.923$ (95% bootstrap CI [0.826, 1.000]) and accuracy = 0.939. The disagreement reflected some boundary cases between Semantic and Transformation, indicating that the fine-grained taxonomy is applied consistently in practice and is suitable for category-level alignment evaluation.

Simulation Sanity Check. We assess simulated queries using semantic similarity to human queries (SemSim; higher is better) and Jensen–Shannon divergence over taxonomy categories (JSDcat; lower is better). Transcript conditioning substantially improves both metrics (SemSim = 0.63, JSDcat = 0.40), with the strongest alignment achieved by combining demographic and transcript information (SemSim = 0.68, JSDcat = 0.31). Demographic-only performs less consistently (SemSim = 0.37, JSDcat = 0.76), and simulations without persona information perform poorly (SemSim = 0.11, JSDcat = 0.87). Therefore, persona conditioning improves alignment with human queries without inducing degenerate outputs.

Profile Stability under Self-consistency. We quantify stability by sampling K profiles per user under the same evidence pack and measuring (i) within-user dispersion (SD/IQR) and (ii) High/Low assignment agreement across samples. Stability varies with inputs: the combined setting Demo+Trans+Logs is the most self-consistent (mean High/Low agreement ≈ 0.90, median within-user SD ≈ 0.05, unstable-trait rate $\approx$ 8%), whereas Demo-only is the least stable (agreement ≈ 0.72, median SD ≈ 0.10, unstable-trait rate $\approx$ 22%). Within Demo+Trans+Logs, *Effort Willingness* and *Strategy Formation* exhibit the highest stability (agreement ≈ 0.93 and ≈ 0.92, respectively; median SD ≈ 0.04 for both), while *System Reliance* is slightly lower (agreement ≈ 0.85, median SD ≈ 0.07), consistent with cases where retrospective self-reports conflict with explicit reliance signals in the interaction logs (e.g., frequent suggestion/autocomplete use despite claimed self-directed querying).

4.3 Inferred Profile Analysis

We cluster users in inferred profiles to identify recurring user types and compare their query formulation behaviors.

Group Differences in Search Behavior. We examine whether inferred user clusters exhibit different query formulation and search behaviors. As shown in Table 4, user groups differ across query-, task-, and session-level characteristics, indicating distinct behavioral profiles and directly addressing **RQ2**.

Group Differences in Demographics. Figure 2 shows clear trait-profile shifts across demographic groupings. Age exhibits the largest separations: younger users are scored higher on *Exploratory Curiosity*, *Effort Willingness*, and *System Reliance*, whereas older users score higher on *Strategy Formation* and slightly higher on *Authority Preference* (with *Evidence Rigor* differing only modestly). Gender differences are more localized: male users show higher *Authority Preference* and *Evidence Rigor*, while female users are slightly higher in *Effort Willingness*; *Exploratory Curiosity*, *System Reliance*, and *Strategy Formation* are comparatively similar across genders. Language proficiency displays a distinct pattern: multilingual users score higher on *Exploratory Curiosity* and *Effort Willingness* (and modestly higher on *System Reliance*), while monolingual users score higher on *Evidence Rigor*; *Authority Preference* and *Strategy Formation* remain aligned between language groups.

5 Discussion

Our study demonstrates that between-user differences are expressed in the first query, prior to system feedback. Under common, controlled information needs, we pinpointed where divergence begins and offer a strategy-level explanation for why downstream effects can appear selective across interfaces and metrics. This extends prior work on demographic and task effects (which largely considered reformulation and session-level measures [39, 61, 64, 75]), as well as extending past system-oriented QV studies (which treated initial variants as interchangeable evaluation inputs rather than strategy-revealing actions [9, 13]). The analysis demonstrates that initial QVs are meaningful signals with consequences for robust evaluation and user-aware measurement as, even under shared information needs, systems may produce heterogeneous outcomes in the face of QVs.

Observed Factors. We examine demographics, tasks, and subjective factors. Prior work considering these factors emphasized reformulation and session-level metrics [28, 42, 71, 72], leaving the initial query underspecified.

Age-related differences were apparent in first-query construction: older participants often encoded structured constraints and operator-like structure, reducing reliance on interaction-driven exploration. Interviews corroborate the pattern: older users described planning criteria upfront, whereas younger users more often delegated intent inference (e.g., "Google usually guesses what you need") or simplified queries to broaden exploration (e.g., "simpler queries yield better results"). Prior work documented age-related differences at later stages of the search session, reporting variations in flexibility, evaluation emphasis, and interaction patterns [14, 19, 32, 39, 74, 82]. These patterns suggest that some later-session age-related differences may be established from the first query. Externalizing constraints in the initial query narrows the

Table 4: Significant effects of inferred profile features on dependent variables at query, task processing, and session levels.

Level	Dependent Variable	Authority Preference	Exploratory Curiosity	Effort Willingness	Evidence Rigor	Strategy Formation	System Reliance
Query	QVs (Count)	–	–	High *	–	Low **	–
	Query Length (Word)	Low *	–	High ***	High **	High ***	–
	Processed Query Length	Low *	–	High ***	High **	High ***	Low *
	Punctuation Count	–	Low ***	–	–	High **	High *
	Novel Terms (Count)	Low **	–	High ***	High **	High ***	Low **
	Search Strategy (First-level)	Low ***	High **	High ***	–	–	–
	Search Strategy (Second-level)	Low ***	High ***	High ***	Low ***	Low ***	High ***
	Jaccard Similarity	–	–	Low *	–	Low **	–
	Semantic Similarity	Low ***	High **	Low ***	–	Low ***	Low ***
	POS Pattern	–	–	–	–	–	–
	Expected Queries	–	High **	High ***	Low **	Low *	High **
	Expected Documents	High **	–	High ***	–	–	High **
	Lexical Density	–	–	Low *	–	Low **	–
	FRE	–	–	–	–	–	–
				Que. High *			
	Query Form	–	–	Top. Low *	–	–	–
	Query Sentiment	–	–	–	–	–	–
Task Proc.	First Click	High **	–	High ***	–	–	–
	Last Click	–	High *	High ***	High ***	High *	–
	Page Submit	Low **	High **	High ***	High ***	High ***	High **
	Click Count	Low **	Low **	–	High **	High *	High **
Session	SERP Dwell Time	–	Low ***	High **	–	High ***	–
	Result Pages Dwell Time	High *	Low **	High **	High ***	High ***	–
	SERP Clicks	–	–	High *	High ***	High *	–
	Follow-up Clicks	–	–	High **	High ***	–	–
	Image Clicks	–	High **	–	–	–	Low *
	Question Clicks	–	–	–	Low **	–	High **
	Product Clicks	–	–	–	High *	–	–
	Tabs Navigated	–	High ***	High ***	–	Low ***	High **

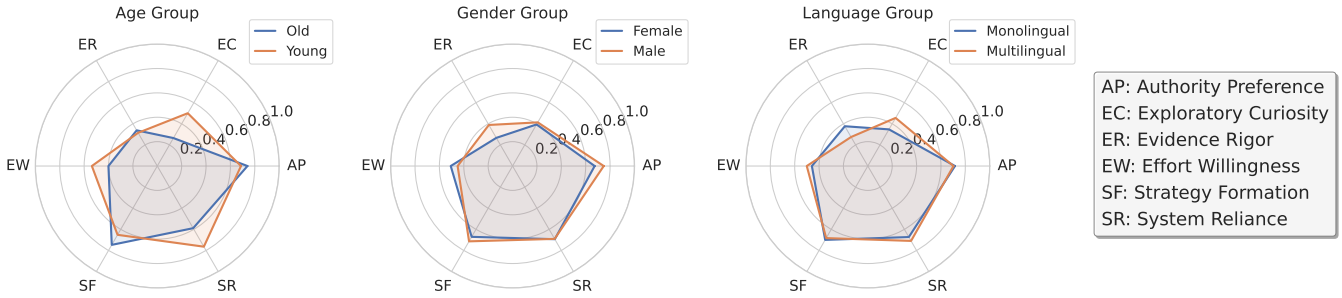


Figure 2: Radar comparison of mean inferred trait scores by demographic group.

retrieved space, shaping subsequent interaction opportunities and associating the first-query strategy with the rest of the session. This motivates us to treat initial query formulation as a first-class stage in user-aware evaluation. By connecting initial formulation to later-session outcomes, we extend prior work and underscore the role of first-query strategies both in interpreting age-related differences and in designing user-aware evaluation metrics.

Prior work on **gender** and **language** effects in search often produced selective or mixed findings [7, 12, 22, 23, 25, 28, 28–30, 33, 34, 53, 57]. Differences primarily manifest as *channel choices*, e.g., question-style formulations and SERP affordances vs. constraint-heavy query encoding. Participant interviews indicate strategic choices rather than differences in retrieval ability. Males described intentionally front-loading constraints to control retrieval scope (e.g., "I include all the criteria needed"; "if the task is subjective, the query will be even longer"), whereas female participants reported different approaches (e.g., "avoid adding too many words"; "I'm always looking for images or visuals to evaluate"). Monolingual participants reported deliberate use of punctuation (e.g., commas, quote marks, "+") to structure intent, demonstrating operator-like strategies rather than artifacts of pre-processing [18, 70].

Task Characteristics. Prior work links task complexity to cognitive effort [9, 35, 40, 59, 78, 80]. Our findings extend this work by showing that cognitively demanding or longer tasks are associated with a greater tendency to front-load constraints and introduce novel terms, whereas simpler or shorter tasks correspond to more

semantically aligned formulations and relatively higher reliance on interaction signals. Subjective factors (familiarity, interest, satisfaction, perceived complexity) are further associated with these choices, indicating that "task difficulty" is not purely objective and that first-query diversity partially reflects perceived rather than nominal task demands [55, 60, 66]. These findings suggest that query structure serves as intentional signals of planning and effort. Recognizing these mechanisms helps avoid over-interpreting demographic correlates as fixed deficits, and emphasizes the importance of modeling early query formulation as a strategy-driven stage in search behavior.

Simulation and Latent Factors. Demographics alone do not explain why users with similar observable attributes produce distinct QVs, and prior user-aware QV generation often relies on synthetic or manually specified personas [3, 83]. By integrating simulation-based inference with observable analyses, we interpret group-level differences as distinct mechanisms inferred from traceable evidence, supporting higher-fidelity simulated QVs. Younger users' interaction-driven exploration corresponds to higher *Exploratory Curiosity* and *Effort Willingness*, whereas older users' formulation-focused strategies correspond to higher *Strategy Formation* and *Authority Preference*. Multilingual users' reliance on interaction and visual content similarly reflects elevated exploratory and effort-related mechanisms. This latent layer complements demographic results by explaining within-group variability and offering a potential explanation for how users with similar observable profiles can arrive at different QVs through different strategies.

The same surface behavior can arise from distinct drivers [49, 54], e.g., active interaction motivated by verification-oriented *Evidence Rigor* versus effort-investment-oriented *Effort Willingness*. Depth–breadth considerations further differentiate Effort Willingness (high-effort, narrow scope) from Evidence Rigor (broader exploration with shallower processing) [43]. Evidence-grounded persona provides a principled way to test whether hypothesized mechanisms are sufficient to reproduce strategy-level properties of human QVs, offering triangulation beyond significance tests on individual observable metrics. We emphasize that these dimensions are intended as *behavioral mechanisms for IR* rather than psychological ground-truth labels. While personality theory can provide an interpretive lens [50, 63], our claims rest on traceable evidence, and should be read as explaining observed search behavior rather than assigning post-hoc psychometric identities.

Implications and Limitations. Our findings extend prior work in three ways. First, because differences emerge from the initial query, robustness-oriented evaluation should treat the *distribution* of plausible QVs as a first-class object rather than assuming a single canonical query [4, 9]. Second, the effort-allocation framing suggests that demographic effects are better interpreted as differences in *where* effort is allocated early (query vs interaction), which helps explain why prior studies report selective or mixed effects across measures and interfaces. Third, by integrating a controlled user study with evidence-based simulation, we demonstrate a tractable path to generating user-aware QVs that are both interpretable and testable, moving beyond demographically sketched or manually specified personas toward empirically grounded mechanisms.

Several limitations remain. Trait validity is necessarily indirect, as we do not establish external psychological ground truth; future statistical modeling could be strengthened by explicitly testing interaction effects in the factorial design and by using mixed-effects models to account for repeated measures and multiple comparisons. Simulation evaluation should further address potential transcript leakage under holdout protocols (e.g., masking task-specific transcript segments for held-out tasks) and report sensitivity to evidence-gating thresholds and sampling choices.

6 Conclusion and Future Work

We asked *how do observable user characteristics and latent search mechanisms shape initial query formulation and early search behavior?* Findings show that behavioral divergence emerges at the outset of search, framing the initial query not as a neutral system input but as an intentional act associated with downstream retrieval and interaction patterns. The main findings with respect to our research questions are as follows:

[RQ1] Observable demographics and task factors shaped initial query formulation and early search interactions. Age showed the strongest and most consistent differences, while gender and language proficiency exhibited more selective differences across query-, task-, and session-level measures; task complexity and length further associated with these patterns.

[RQ2] Our evidence-grounded simulation framework inferred stable, interpretable latent mechanisms from interaction logs and post-task rationales without assuming ground-truth psychological labels. Persona conditioning improved alignment to human query sequences over task-only baselines, and transcript-derived evidence contributed more than demographics alone; the resulting profiles remain measurable, auditable, and internally consistent.

[RQ3] The inferred six-dimensional trait space complements demographic analyses by enabling compact group-level comparisons and explaining within-group variability in exploration, effort allocation, and evidence orientation. The findings indicate that both observable and latent user factors shape search behavior, motivating user-aware system design and robustness-oriented evaluation that account for demographics, task context, and inferred mechanisms.

This study used a methodological framework for studies of query variants, enabling the joint analysis of observable user characteristics and latent search mechanisms. Key factors associated with query formulation and early search behavior were identified. Future work can explore interactions among demographic variables and individual differences, enabling more fine-grained and potentially real-time, profile-informed adaptations. Extending to diverse tasks, larger populations, and more naturalistic search settings, spanning varied devices and cognitive loads, can further capture behavioral complexity and support efficient, fair, and adaptive search systems.

7 Acknowledgments

We acknowledge the support of the Saudi Arabian Cultural Mission in Australia and the Ministry of Education of Saudi Arabia. This research was conducted at the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) and funded by the Australian Research Council (CE200100005). We thank RACE (RMIT Advanced Cloud Ecosystem) for providing computing resources.

References

[1] Nuha Abu Onq. 2023. How do Human and Contextual Factors Affect the Way People Formulate Queries?. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 499–503.

[2] Nuha Abu Onq, Mark Sanderson, and Falk Scholer. 2025. Classifying term variants in query formulation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2395–2406.

[3] Marwah Alaofi, Nicola Ferro, Paul Thomas, Falk Scholer, and Mark Sanderson. 2025. Demographically-Inspired Query Variants Using an LLM. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. 390–400.

[4] 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 *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2850–2862. https://doi.org/10.1145/3477495.3531711

[5] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 1869–1873. https://doi.org/10.1145/3539618.3591960

[6] Abhijit Anand, Jurek Leonhardt, Venktesh V, and Avishek Anand. 2024. Understanding the User: An Intent-Based Ranking Dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, Edoardo Serra and Francesca Spezzano (Eds.). ACM, 5323–5327. https://doi.org/10.1145/3627673.3679166

[7] Manon Arcand and Jacques Nantel. 2012. Uncovering the Nature of Information Processing of Men and Women Online: The Comparison of Two Models Using the Think-Aloud Method. *J. Theor. Appl. Electron. Commer. Res.* 7, 2 (2012), 106–120. https://doi.org/10.4067/S0718-18762012000200010

[8] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 725–728. https://doi.org/10.1145/2911451.2914671

[9] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 395–404. https://doi.org/10.1145/3077136.3080839

[10] Krisztian Balog and ChengXiang Zhai. 2023. User simulation for evaluating information access systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 302–305.

[11] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2012. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 105–114.

[12] Emad Bataineh and Bilal Al-Bataineh. 2014. An analysis study on how female college students view the web search results using eye tracking methodology. In *Proceedings of the International Conference on Human-Computer Interaction, Crete, Greece*. 22–27.

[13] Rodger Benham, Joel M. Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting Search Performance Using Query Variations. *ACM Trans. Inf. Syst.* 37, 4 (2019), 41:1–41:25. https://doi.org/10.1145/3345001

[14] Jennifer C. Romano Bergstrom, Erica L. Olmsted-Hawala, and Matt E. Jans. 2013. Age-Related Differences in Eye Tracking and Usability Performance: Website Usability for Older Adults. *Int. J. Hum. Comput. Interact.* 29, 8 (2013), 541–548. https://doi.org/10.1080/10447318.2012.728493

[15] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: social data meets search queries. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo Baeza-Yates, and Sue B. Moon (Eds.). International World Wide Web Conferences Steering Committee / ACM, 131–140. https://doi.org/10.1145/2488388.2488401

[16] Toine Bogers, Maria Gäde, Mark Hall, and Mette Skov. 2016. Analyzing the influence of language proficiency on interactive book search behavior. *IConference 2016 proceedings* (2016).

[17] Timo Breuer. 2024. Data Fusion of Synthetic Query Variants With Generative Large Language Models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 274–279.

[18] Gracinda Carvalho, David Martins de Matos, and Vítor Rocio. 2007. Document retrieval for question answering: a quantitative evaluation of text preprocessing. In *Proceedings of the First Ph.D. Workshop in CIKM, PIKM 2007, Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 9, 2007*, Aparna S. Varde and Jian Pei (Eds.). ACM, 125–130. https://doi.org/10.1145/1316874.1316894

[19] Aline Chevalier, Aurélie Dommes, and Jean Claude Marquié. 2015. Strategy and accuracy during information search on the Web: Effects of age and complexity of the search questions. *Comput. Hum. Behav.* 53 (2015), 305–315. https://doi.org/10.1016/J.CHB.2015.07.017

[20] Jessie Chin, Evan Anderson, Chieh-Li Chin, and Wai-Tat Fu. 2015. Age differences in information search: An exploration-exploitation tradeoff model. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*, David C. Noelle, Rick Dale, Anne S. Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn D. Jennings, and Paul P. Maglio (Eds.). cognitivesciencesociety.org. https://escholarship.org/uc/item/4w33c6wj

[21] Peng Chu, Eszter Józsa, Anita Komlodi, and Károly Hercegfi. 2012. An exploratory study on search behavior in different languages. In *Information Interaction in Context: 2012, IIix'12, Nijmegen, The Netherlands, August 21-24, 2012*, Jaap Kamps, Wessel Kraaij, and Norbert Fuhr (Eds.). ACM, 318–321. https://doi.org/10.1145/2362724.2362784

[22] Peng Chu, Anita Komlodi, and Gyöngyi Rózsa. 2015. Online search in english as a non-native language. In *Information Science with Impact: Research in and for the Community - Proceedings of the 78th ASIS&T Annual Meeting, ASIST 2015, St. Louis, Missouri, Missouri, USA, October 6-10, 2015 (Proc. Assoc. Inf. Sci. Technol., Vol. 52)*. Wiley, 1–9. https://doi.org/10.1002/PRA2.2015.145052010040

[23] Janne Chung and Gary Monroe. 1998. Gender differences in information processing: an empirical test of the hypothesis-confirming strategy in an audit context. *Accounting & Finance* 38, 2 (1998), 265–279.

[24] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2022. Topic Difficulty: Collection and Query Formulation Effects. *ACM Trans. Inf. Syst.* 40, 1 (2022), 19:1–19:36. https://doi.org/10.1145/3470563

[25] William K Darley and Robert E Smith. 1995. Gender differences in information processing strategies: An empirical test of the selectivity model in advertising response. *Journal of advertising* 24, 1 (1995), 41–56.

[26] Aurelie Dommes, Aline Chevalier, and Sarah Lia. 2011. The role of cognitive flexibility and vocabulary abilities of younger and older users in searching for information on the web. *Applied Cognitive Psychology* 25, 5 (2011), 717–726.

[27] Peter G. Fairweather. 2008. How older and younger adults differ in their approach to problem solving on a complex website. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2008, Halifax, Nova Scotia, Canada, October 13-15, 2008*, Simon Harper and Armando Barreto (Eds.). ACM, 67–72. https://doi.org/10.1145/1414471.1414485

[28] Hengyi Fu. 2017. Query Reformulation Patterns of Mixed Language Queries in Different Search Intents. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel (Eds.). ACM, 249–252. https://doi.org/10.1145/3020165.3022126

[29] Hengyi Fu and Shuheng Wu. 2014. Analyzing chinese-english mixed language queries in a web search engine. In *Connecting Collections, Cultures, and Communities - Proceedings of the 77th ASIS&T Annual Meeting, ASIST 2014, Seattle, WA, USA, October 31 - November 5, 2014 (Proc. Assoc. Inf. Sci. Technol., Vol. 51)*. Wiley, 1–5. https://doi.org/10.1002/MEET.2014.14505101114

[30] M. Rami Ghorab, Séamus Lawless, Alexander O'Connor, Dong Zhou, and Vincent Wade. 2013. Multilingual vs. Monolingual User Models for Personalized Multilingual Information Retrieval. In *User Modeling, Adaptation, and Personalization - 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013, Proceedings (Lecture Notes in Computer Science, Vol. 7899)*, Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro (Eds.). Springer, 356–358. https://doi.org/10.1007/978-3-642-38844-6_38

[31] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric Analysis of Bloggers' Age and Gender. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*, Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng (Eds.). The AAAI Press. http://aaai.org/ocs/index.php/ICWSM/09/paper/view/208

[32] Jacek Gwizdka and Dania Bilal. 2017. Analysis of Children's Queries and Click Behavior on Ranked Results and Their Thought Processes in Google Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel (Eds.). ACM, 377–380. https://doi.org/10.1145/3020165.3022157

[33] Yoo Jin Ha. 2008. *Accessing and using multilanguage information by users searching in different information retrieval systems.* Rutgers The State University of New Jersey, School of Graduate Studies.

[34] Scott A. Hale. 2014. Multilinguals and Wikipedia editing. In *ACM Web Science Conference, WebSci '14, Bloomington, IN, USA, June 23-26, 2014*, Filippo Menczer, Jim Hendler, William H. Dutton, Markus Strohmaier, Ciro Cattuto, and Eric T. Meyer (Eds.). ACM, 99–108. https://doi.org/10.1145/2615569.2615684

[35] Jiyin He and Emine Yilmaz. 2017. User Behaviour and Task Characteristics: A Field Study of Daily Information Behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel (Eds.). ACM, 67–76. https://doi.org/10.1145/3020165.3020188

[36] Daniel Hienert and Maria Lusky. 2017. Where Do All These Search Terms Come From? - Two Experiments in Domain-Specific Search. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings (Lecture Notes in Computer Science, Vol. 10193)*, Joemon M. Jose, Claudia Hauff, Ismail Sengör Altingövde, Dawei Song, Dyaa Albakour, Stuart N. K. Watt, and John Tait (Eds.). 15–26. https://doi.org/10.1007/978-3-319-56608-5_2

[37] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[38] Maria Kanwal, Umar Burki, Raza Ali, and Robert Dahlstrom. 2022. Systematic review of gender differences and similarities in online consumers' shopping behavior. *Journal of Consumer Marketing* 39, 1 (2022), 29–43.

[39] Saraschandra Karanam and Herre van Oostendorp. 2016. Age-related Differences in the Content of Search Queries when Reformulating. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM, 5720–5730. https://doi.org/10.1145/2858036.2858444

[40] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-Ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR 2015, Northampton, Massachusetts, USA, September 27-30, 2015*, James Allan, W. Bruce Croft, Arjen P. de Vries, and Chengxiang Zhai (Eds.). ACM, 101–110. https://doi.org/10.1145/2808194.2809465

[41] Dae-Young Kim, Xinran Y Lehto, and Alastair M Morrison. 2007. Gender differences in online travel information search: Implications for marketing communications on the internet. *Tourism management* 28, 2 (2007), 423–433.

[42] Khamsum Kinley, Dian Tjondronegoro, Helen Partridge, and Sylvia L. Edwards. 2012. Human-computer interaction: the impact of users' cognitive styles on query reformulation behaviour during web searching. In *The 24th Australian Computer-Human Interaction Conference, OzCHI '12, Melbourne, VIC, Australia - November 26 - 30, 2012*, Vivienne Farrell, Graham Farrell, Caslon Chua, Weidong Huang, Rajesh Vasa, and Clinton Woodward (Eds.). ACM, 299–307. https://doi.org/10.1145/2414536.2414586

[43] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. 2004. Depth-and breadth-first processing of search result lists. In *CHI'04 extended abstracts on Human factors in computing systems*. 1539–1539.

[44] Anett Kralisch and Bettina Berendt. 2005. Language-sensitive search behaviour and the role of domain knowledge. *New Rev. Hypermedia Multim.* 11, 2 (2005), 221–246. https://doi.org/10.1080/13614560500402775

[45] Andrew Large, Jamshid Beheshti, and Tarjin Rahman. 2002. Gender differences in collaborative Web searching behavior: an elementary school study. *Inf. Process. Manag.* 38, 3 (2002), 427–443. https://doi.org/10.1016/S0306-4573(01)00034-6

[46] Binsheng Liu, Nick Craswell, Xiaolu Lu, Oren Kurland, and J. Shane Culpepper. 2019. A Comparative Analysis of Human and Automatic Query Variants. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019*, Yi Fang, Yi Zhang, James Allan, Krisztian Balog, Ben Carterette, and Jiafeng Guo (Eds.). ACM, 47–50. https://doi.org/10.1145/3341981.3344223

[47] Lori Lorigo, Bing Pan, Helene Hembrooke, Thorsten Joachims, Laura A. Granka, and Geri Gay. 2006. The influence of task and gender on search and evaluation behavior using Google. *Inf. Process. Manag.* 42, 4 (2006), 1123–1131. https://doi.org/10.1016/J.IPM.2005.10.001

[48] Ryan Lowe and Ben Steichen. 2017. Multilingual Search User Behaviors - Exploring Multilingual Querying and Result Selection Through Crowdsourcing. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP 2017, Bratislava, Slovakia, July 09 - 12, 2017*, Mária Bieliková, Eelco Herder, Federica Cena, and Michel C. Desmarais (Eds.). ACM, 303–307. https://doi.org/10.1145/3079628.3079702

[49] Chenglong Ma, Yongli Ren, Pablo Castells, and Mark Sanderson. 2022. NEST: simulating pandemic-like events for collaborative filtering by modeling user needs evolution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1430–1440.

[50] Chenglong Ma, Ziqi Xu, Yongli Ren, Danula Hettiachchi, and Jeffrey Chan. 2025. PUB: an LLM-enhanced personality-driven user behaviour simulator for recommender system evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2690–2694.

[51] Joel M. Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A Large English News Corpus. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 3077–3084. https://doi.org/10.1145/3340531.3412762

[52] Parinaz Maghferat and Wolfgang G. Stock. 2010. Gender-specific information search behavior. *Webology* 7, 2 (2010). http://www.webology.org/2010/v7n2/a80.html

[53] Mari Carmen Marcos, Ruth Olimpia García-Gavilanes, Emad Bataineh, and Lara Pasarin. 2013. Using eye tracking to identify cultural differences in information seeking behavior. (2013).

[54] AH Maslow. 1943. A theory of human motivation. *Psychological Review google schola* 2 (1943), 21–28.

[55] Heather L. O'Brien, Jaime Arguello, and Robert Capra. 2020. An empirical study of interest, task complexity, and search behaviour on user engagement. *Inf. Process. Manag.* 57, 3 (2020), 102226. https://doi.org/10.1016/J.IPM.2020.102226

[56] Ed O'Donnell and Eric N Johnson. 2001. The effects of auditor gender and task complexity on information processing efficiency. *International journal of auditing* 5, 2 (2001), 91–105.

[57] Veronika Papyrina. 2015. Men and women watching and reading: Gender and information processing opportunity effects in advertising. *Journal of Marketing Communications* 21, 2 (2015), 125–143.

[58] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 397–412. https://doi.org/10.1007/978-3-030-99736-6_27

[59] Yan Qu and George W. Furnas. 2008. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Inf. Process. Manag.* 44, 2 (2008), 534–555. https://doi.org/10.1016/J.IPM.2007.09.006

[60] Tara L Queen, Thomas M Hess, Gilda E Ennis, Keith Dowd, and Daniel Grühn. 2012. Information search and decision making: effects of age and complexity on strategy use. *Psychology and aging* 27, 4 (2012), 817.

[61] Amifa Raj, Bhaskar Mitra, Nick Craswell, and Michael D. Ekstrand. 2023. Patterns of Gender-Specializing Query Reformulation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2241–2245. https://doi.org/10.1145/3539618.3592034

[62] Haywantee Ramkissoon and Robin Nunkoo. 2012. More than just biological sex differences: Examining the structural relationship between gender identity and information search behavior. *Journal of Hospitality & Tourism Research* 36, 2 (2012), 191–215.

[63] Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin* 28, 6 (2002), 789–801.

[64] Sophie A. Rutter, Nigel Ford, and Paul D. Clough. 2015. How do children reformulate their search queries? *Inf. Res.* 20, 1 (2015). http://www.informationr.net/ir/20-1/isic2/isic31.html

[65] Mylene Sanchiz, Franck Amadieu, Wai Tat Fu, and Aline Chevalier. 2019. Does pre-activating domain knowledge foster elaborated online information search strategies? Comparisons between young and old web user adults. *Applied ergonomics* 75 (2019), 201–213.

[66] Mylène Sanchiz, Aline Chevalier, and Franck Amadieu. 2017. How do older and young adults start searching for information? Impact of age, domain knowledge and problem complexity on the different steps of information searching. *Comput. Hum. Behav.* 72 (2017), 67–78. https://doi.org/10.1016/J.CHB.2017.02.038

[67] Ivan Sekulić, Lili Lu, Navdeep Singh Bedi, and Fabio Crestani. 2024. Simulating Conversational Search Users with Parameterized Behavior. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 72–81.

[68] 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 (2008), 3:1–3:25. https://doi.org/10.1145/1352782.1352785

[69] Georg Singer, Ulrich Norbisrath, and Dirk Lewandowski. 2012. Impact of Gender and Age on performing Search Tasks Online. In *Mensch & Computer 2012: interaktiv informiert - allgegenwärtig und allumfassend!? Konstanz, Germany, September 9-12, 2012*, Harald Reiterer and Oliver Deussen (Eds.). Oldenbourg Verlag, 23–32. https://dl.gi.de/handle/20.500.12116/7800

[70] Ellen Souza, Douglas Vitório, Gyovana Moriyama, Luiz Santos, Lucas Martins, Mariana Souza, Márcio Fonseca F. da Silva, Nádia Félix, André C. P. L. F. de Carvalho, Hidelberg Oliveira Albuquerque, and Adriano L. I. Oliveira. 2021. An Information Retrieval Pipeline for Legislative Documents from the Brazilian Chamber of Deputies. In *Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021 (Frontiers in Artificial Intelligence and Applications, Vol. 346)*, Erich Schweighofer (Ed.). IOS Press, 119–126. https://doi.org/10.3233/FAIA210326

[71] Ben Steichen, M. Rami Ghorab, Alexander O'Connor, Séamus Lawless, and Vincent Wade. 2014. Towards Personalized Multilingual Information Access - Exploring the Browsing and Search Behavior of Multilingual Users. In *User Modeling*,

Adaptation, and Personalization - 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings (Lecture Notes in Computer Science, Vol. 8538), Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben (Eds.). Springer, 435–446. https://doi.org/10.1007/978-3-319-08786-3_39

[72] Ben Steichen and Ryan Lowe. 2021. How do multilingual users search? An investigation of query and result list language choices. *J. Assoc. Inf. Sci. Technol.* 72, 6 (2021), 759–776. https://doi.org/10.1002/ASI.24443

[73] Aideen J. Stronge, Wendy A. Rogers, and Arthur D. Fisk. 2006. Web-Based Information Search and Retrieval: Effects of Strategy Use and Age on Search Success. *Hum. Factors* 48, 3 (2006), 434–446. https://doi.org/10.1518/001872006778606804

[74] Sergio Duarte Torres, Ingmar Weber, and Djoerd Hiemstra. 2014. Analysis of Search and Browsing Behavior of Young Users on the Web. *ACM Trans. Web* 8, 2 (2014), 7:1–7:54. https://doi.org/10.1145/2555595

[75] Herre van Oostendorp, Tijs Ditvoorst, and Justin van Doorn. 2018. Feedback on the Semantic Relevance of Search Queries. In *Proceedings of the 36th European Conference on Cognitive Ergonomics, Utrecht, The Netherlands, September 05-07, 2018*. ACM, 25:1–25:4. https://doi.org/10.1145/3232078.3232087

[76] Viswanath Venkatesh and Michael G. Morris. 2000. Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. *MIS Q.* 24, 1 (2000), 115–139. http://misq.org/why-don-t-men-ever-stop-to-ask-for-directions-gender-social-influence-and-their-role-in-technology-acceptance-and-usage-behavior.html

[77] Ingmar Weber and Alejandro Jaimes. 2011. Who uses web search for what: and how. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, Irwin King, Wolfgang Nejdl, and Hang Li (Eds.). ACM, 15–24. https://doi.org/10.1145/1935826.1935839

[78] Barbara M. Wildemuth, Diane Kelly, Emma Boettcher, Erin Moore, and Gergana Dimitrova. 2018. Examining the impact of domain and cognitive complexity on query formulation and reformulation. *Inf. Process. Manag.* 54, 3 (2018), 433–450. https://doi.org/10.1016/j.ipm.2018.01.009

[79] Chuhan Wu, Fangzhao Wu, Junxin Liu, Shaojian He, Yongfeng Huang, and Xing Xie. 2019. Neural Demographic Prediction using Search Query. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 654–662. https://doi.org/10.1145/3289600.3291034

[80] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, tanning beds, tattoos and NASCAR: evaluation of search tasks with varying levels of cognitive complexity. In *Information Interaction in Context: 2012, IIix'12, Nijmegen, The Netherlands, August 21-24, 2012*, Jaap Kamps, Wessel Kraaij, and Norbert Fuhr (Eds.). ACM, 254–257. https://doi.org/10.1145/2362724.2362768

[81] Elad Yom-Tov. 2019. Demographic differences in search engine use with implications for cohort selection. *Inf. Retr. J.* 22, 6 (2019), 570–580. https://doi.org/10.1007/S10791-018-09349-2

[82] Robert J. Youmans, Brooke G. Bellows, Christian A. Gonzalez, Brittany Sarbone, and Ivonne J. Figueroa. 2013. Designing for the Wisdom of Elders: Age Related Differences in Online Search Strategies. In *Universal Access in Human-Computer Interaction. User and Context Diversity - 7th International Conference, UAHCI 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 8010)*, Constantine Stephanidis and Margherita Antona (Eds.). Springer, 240–249. https://doi.org/10.1007/978-3-642-39191-0_27

[83] Oleg Zendel, Sara Fahad Dawood Al Lawati, Lida Rashidi, Falk Scholer, and Mark Sanderson. 2025. A Comparative Analysis of Linguistic and Retrieval Diversity in LLM-Generated Search Queries. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, Meeyoung Cha, Chanyoung Park, Noseong Park, Carl Yang, Senjuti Basu Roy, Jessie Li, Jaap Kamps, Kijung Shin, Bryan Hooi, and Lifang He (Eds.). ACM, 4014–4023. https://doi.org/10.1145/3746252.3761382

[84] Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2687–2692.

[85] Pengyi Zhang, Chang Liu, and Preben Hansen. 2016. I Need More Time!: The Influence of Native Language on Search Behavior and Experience. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016 (CEUR Workshop Proceedings, Vol. 1609)*, Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald (Eds.). CEUR-WS.org, 1166–1182. https://ceur-ws.org/Vol-1609/16091166.pdf

[86] Mingming Zhou. 2014. Gender difference in web search perceptions and behavior: Does it vary by task performance? *Comput. Educ.* 78 (2014), 174–184. https://doi.org/10.1016/J.COMPEDU.2014.06.005