RESEARCH-ARTICLE

Comparing Traditional and LLM-based Search for Image Geolocation

ALBATOOL WAZZAN, Temple University, Philadelphia, PA, United States

STEPHEN MACNEIL, Temple University, Philadelphia, PA, United States

RICHARD M SOUVENIR, Temple University, Philadelphia, PA, United States

Published: 10 March 2024

CHIIR '24: 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval
March 10 - 14, 2024
Sheffield, United Kingdom

CHIIR '24: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (March 2024)
https://doi.org/10.1145/3627508.3638305
ISBN: 9798400704345

Comparing Traditional and LLM-based Search for Image Geolocation

Albatool Wazzan
Dept of Computer & Info Sciences
Temple University
Philadelphia, USA
albatool.wazzan@temple.edu

Stephen MacNeil
Dept of Computer & Info Sciences
Temple University
Philadelphia, USA
stephen.macneil@temple.edu

Richard Souvenir
Dept of Computer & Info Sciences
Temple University
Philadelphia, USA
souvenir@temple.edu

ABSTRACT

Web search engines have long served as indispensable tools for information retrieval; user behavior and query formulation strategies have been well studied. The introduction of search engines powered by large language models (LLMs) suggested more conversational search and new types of query strategies. In this paper, we compare traditional and LLM-based search for the task of image geolocation, i.e., determining the location where an image was captured. Our work examines user interactions, with a particular focus on query formulation strategies. In our study, 60 participants were assigned either traditional or LLM-based search engines as assistants for geolocation. Participants using traditional search more accurately predicted the location of the image compared to those using the LLM-based search. Distinct strategies emerged between users depending on the type of assistant. Participants using the LLM-based search issued longer, more natural language queries, but had shorter search sessions. When reformulating their search queries, traditional search participants tended to add more terms to their initial queries, whereas participants using the LLM-based search consistently rephrased their initial queries.

ACM Reference Format:
Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing Traditional and LLM-based Search for Image Geolocation. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3627508.3638305

1 INTRODUCTION

For decades, web search engines have served as the de facto reference tool for a wide range of tasks. In fact, it has been demonstrated that humans have been trained to optimize keyword-based searching using query formulations not typically used in natural language [24]. Advancements in artificial intelligence (AI) have driven the emergence of large language models (LLM), such as BERT [9], GPT-3 [5], and their successors. These models have served as the foundation for numerous applications, ranging from text generation, translation, to question answering, multi-step reasoning, and complex problem solving [43]. Recently, these tools have been combined with web search to enable a new mode of LLM-powered conversational search. Unlike keyword-based search, this integration allows users to engage in a natural, interactive conversation with the LLM-powered search engine, as if they were interacting with a knowledgeable assistant. This conversational mode has the potential to improve user experiences across various domains. While previous work has explored how people engage in sense-making and constructing mental models of traditional search engines [40], the adaptation of these models to LLM-based search remains open to inquiry.

To compare traditional and LLM-based search, we consider the task of image geolocation – identifying the location in which an image was captured, an important task with applications in forensics, law enforcement, and journalism. This task has historically been performed by expert image analysts, using increasingly sophisticated reference tools as they became available. Fully automated computer vision approaches [12, 26, 45] have been developed; these approaches typically rely on the visual similarity between the query image and a previously-processed training image and tend to work best when landmarks or other unique features are visible. In the general case, accurately localizing images can be challenging. Even with the assistance of a search engine, users not only need to identify visual clues, but understand them well enough to translate into a search query. Because geolocation is a task that requires investigation, in that analysts must collect sometimes disparate clues to uncover the origin of the image, it can be expected that users will formulate multiple queries as they seek to retrieve information about these clues. This task takes advantage of both the lookup abilities of a search engine and the contextual knowledge from humans, making it a compelling task to evaluate how users adapt their query formulation strategies.

We conducted a between-subjects study with 60 participants randomly assigned to use either traditional or LLM-based search to aid in image geolocation in order to address the following research questions:

RQ1 How does the use of an LLM-based search tool versus a traditional search tool impact participants' performance in geolocation tasks?

RQ2 How do participants adapt their query formulation strategies when using LLM-based search compared to traditional search for image geolocation?

RQ3 What are the key challenges encountered by participants when using LLM-based search for image geolocation?

Our results indicate that participants using traditional search outperformed those using LLM-based search in terms of accurate image

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638305

geolocation. This outcome can be explained by our qualitative findings, where participants reported challenges formulating queries when interacting with the LLM-based search engine. LLM-based search users issued longer, more conversational queries within shorter search sessions. Participants using the traditional search engine tended to extend their initial queries with additional terms when reformulating, while those utilizing the LLM-based search consistently rephrased their initial queries.

2 RELATED WORK

Search engines have evolved into indispensable tools that influence how information is accessed and problems are solved [11]. However, effectively communicating the user's search intent has been a persistent challenge. Much work has been dedicated toward understanding web search query formulation patterns [3, 18, 30] and investigating how users adapt their queries and reformulation strategies in efforts to uncover search intent [6, 20, 42]. These strategies can be domain-specific. For instance, for health-related information, Zuccon showed that search results were less helpful when users issued complex queries describing their symptoms rather than using medical terminology [48]. In the educational setting, students heavily rely on search engines for academic purposes [17, 36]. However, it has been shown that a substantial portion of academic search sessions result in null queries, when individuals use vague or complex terms resulting in empty search results and obstructing users from achieving their intended search objectives [25]. Recognizing and understanding these challenges related to user behavior and query formulation strategies can enhance the overall search and retrieval experience. Our work builds upon existing research in web search and query formulation and extends the analysis to LLM-based search for the task of image geolocation.

2.1 LLM-based Search Analysis

LLMs are trained on large amounts of text corpra, and their effectiveness in various applications hinges on the ability to query them effectively [1]. In efforts to optimize LLMs for retrieval tasks, several works have investigated the process of querying LLMs for specific information. Jiang [21] highlighted the consequences of poorly written prompts, yielding failed retrieval results and proposed a method that used multiple automated paraphrases of the query and an aggregation scheme, mirroring how humans often rephrase their queries and provide additional context to make them more informative. Similarly, the work of Petroni [32] examined enhancing the LLM retrieval by augmenting queries with relevant context and demonstrated improved performance on various LLMs on factuality tests.

When humans seek information, they often clarify their queries with examples to obtain better results. To mimic this behavior, Brown employed few-shot learning, which involves conditioning the LLM on the task description and just a few examples, and found that this "in-context learning" works best with larger language models [5]. A more recent effort argued that language models do not learn tasks during runtime from few-shot examples, but locate tasks within the model's preexisting knowledge; this paper proposes 0-shot prompts, which uses an alternative query with different phrasings to provide additional task descriptions [35]. Wei introduces chain-of-thought, which aims to replicate the human thought process when addressing complex problems [44]. These efforts aimed at enhancing the LLM retrieval, but do not address the challenges faced by non-expert users when querying LLMs. Recent work [47] involves non-experts issuing prompts for LLM-based chatbots and found that struggles in formulating effective prompts resemble issues observed in end-user programming and interactive machine learning systems. Their work emphasizes the need for further research in LLMs and prompt-literacy, specifically for non-expert users. Our work focuses on this challenge for the multifaceted task of image geolocation.

2.2 Image Geolocation

Image geolocation is a widely-studied task. One effort used a carefully constructed dataset to investigate the types of clues and strategies users employ for image geolocation [29]. Several efforts have addressed the labor-intensive nature of image geolocation by incorporating crowdsourcing to improve location identification. One study introduced a diagramming technique involving visual representations from a bird's-eye or satellite perspective, which allowed novice crowd workers to collaborate with experts [23]. In a follow-up study, the authors introduced GroundTruth, a system that enhanced image geolocation accuracy through shared representations for crowd-augmented expert work [41].

Other studies have explored how to improve the accuracy of non-expert workers in image geolocation tasks. One project explicitly instructed novice users to follow a three-step workflow inferred from expert strategies: collecting image-related clues, deriving potential coordinates based on these clues, and identifying the image location on a map [22]. Another method [34] introduced a crowdsourcing platform that leverages existing data mining methods to estimate photo and video locations from social media, then used crowdsourcing for verification.

In our approach, we focus on how participants articulate visual clues into search queries, and whether those query strategies differ depending on the type of search tool available.

3 METHODS

We conducted a between-subjects user study involving 60 participants. In this section, we describe the experimental platform, recruitment of participants, task design, and measures.

3.1 Experimental Platform

Image geolocation has been well-studied due, in part, to the popularity of gamified versions of the task. The most well-known version is GeoGuessr; others include GeoGuess, Geotastic, and City Guesser. The objective of these games is to predict the correct location on map given an image, video, or other information and points are accumulated based on speed and/or accuracy. These games can serve as useful platforms for evaluating a wide variety of cognitive tasks. In this study, we use GeoGuess [4], an open-source image geolocation game. Users are presented a Google Map StreetView image and have two minutes to guess the location by dropping a pin on a world map. Users can navigate using the StreetView interface to virtually zoom and move through the scene. Up to 5,000 points can be earned based on proximity of the prediction to the actual

Figure 1: Initial viewpoints (with the location indicated) of the six rounds in the experiment.

location. In the instructions for the game, users are encouraged to seek out clues that may be useful for localization.

3.2 Participants

Participants, 18 years of age or older who could read and understand English, were recruited on a university campus. Our sample consisted of 60 participants whose mean age was 25 (SD = 4.95) with a diverse range of academic backgrounds, majoring in accounting, biology, computer science, systems engineering, physics, chemistry, global studies, and marketing. The IRB approved study was carried out by two members of the research team over the span of three weeks. All participants received a $5 gift card to a coffee shop and were informed of the task description, duration, compensation, and their right to forfeit at anytime before participating.

3.3 Task Design

Users were provided a dual-monitor setup, with the geolocation task on one screen and the search engine on the other. Microsoft Bing served as the traditional search engine and Microsoft Bing Chat, which is powered by ChatGPT, served as the LLM-based search engine.

We followed a between-subjects study design, where each participant was randomly assigned to either the (Traditional) Search or the LLM condition. The experiment consisted of six geolocation tasks (shown in Figure 1), which were intended to vary in difficulty. Participants were provided instructions to only use the provided search engine and not perform image-based search. Participants were asked to confirm their understanding of the instructions by clicking on a (I understand) button. After the instructions, participants watched a short instructional video on how to use the geolocation interface.

For each round, the participant had two minutes to provide a guess. They could consult the search tool as often as they needed, given the time constraint. Upon completion of the six rounds, participants were invited to fill out a post-study survey with questions about familiarity with image geolocation, traditional or LLM-based search, attitudes toward artificial intelligence, and a set of open-ended questions for additional feedback. The entire experiment was designed to be completed in approximately 15 minutes per participant.

Table 1: Geolocation performance on the six round experiment.

	Estimate	Std. Error	t-value	p-value
(Intercept)	2678.5	167.4	15.999	<2e-16 ***
Condition(Search)	501.3	243.5	2.059	0.0414 *

3.4 Measures

The primary dependent variable in this study is performance, measured by the points earned by each participant per round. The score ranged from 0 to 5000; the maximum score was obtained when the prediction was within a few kilometers of the actual location. The primary independent variable, type of search, was modeled as a fixed effect in our linear mixed-effects model. Each participant played the six rounds in the same order. The round number, which correlated with difficulty, was modeled as a random effect. For each participant, we maintained an event log of timestamped actions that included switching between web search and geolocation. Additionally, we recorded the search queries.

4 ANALYSIS & RESULTS

We excluded the five participants that did not engage with the search engine for any of the rounds, which left 29 for the Search condition and 26 for the LLM condition.

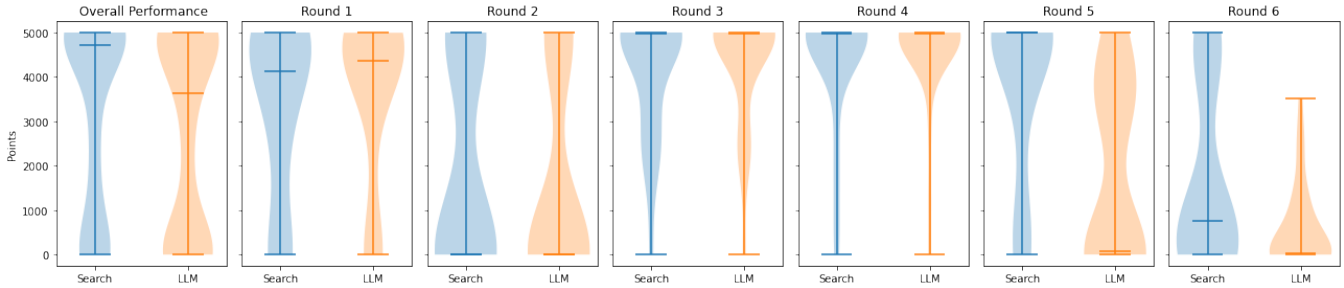


Figure 2: Performance distribution (points) comparison between Search and LLM conditions on average (left) and per round.

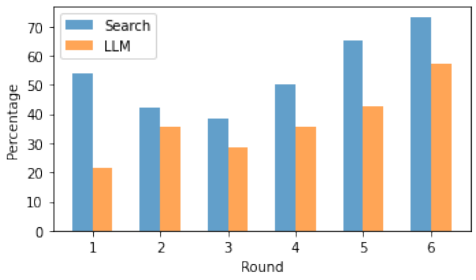


Figure 3: Percentage of multi-query rounds for Search and LLM conditions

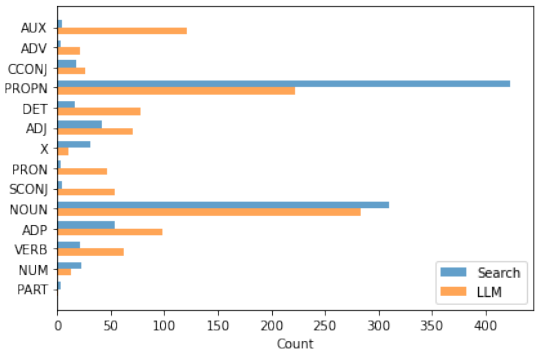


Figure 4: Comparison of part-of-speech tag counts between Search and LLM conditions

For the Search condition, the mean performance score was 3189, with a median of 4712 and an IQR of 4262.25. For the LLM condition, the mean performance score was 2725, with a median of 3637.5 and an IQR of 4952. Figure 2 shows the distribution of scores across the two conditions by average and across the six rounds.

Overall, participants in the Search condition outperformed those in the LLM condition. For rounds 1-4, the performance was similar for both conditions, with both groups finding round 2 challenging. In the final two rounds, participants across both conditions performed poorly, with those in the LLM group performing notably worse. We use a Linear mixed-effects Model (LMM) to evaluate the difference in performance between the groups. The results are shown in Table 1. There was a significant difference in performance between the two conditions ($p = 0.0414$).

4.1 Query Formulation Patterns

Query formulation is a fundamental tool in the analysis of search behavior. Here, we investigate four key query formulation metrics, comparing their differences across the two conditions.

4.1.1 Number of Queries. We examined the number of queries per round. Participants in the Search condition issued an average of 1.98 queries, whereas those in the LLM condition issued an average of 1.04 queries. A Chi-Square test showed this association to be significant ($\chi^2(6) = 19.71, p = 0.003$); users in the Search condition issued more queries.

We computed the percentage of rounds in which participants issued more than one query. As shown in Figure 3, participants in both conditions issued more queries as the task increased in difficulty. Participants in the Search condition favored issuing more queries, starting at around 55% and reaching 70% by the end of the task. In the LLM condition, participants issued more queries at a lower rate starting at 20%, but reaching 55% by the last round.

4.1.2 Query Length. Query length, the average number of terms in each query, can provide valuable insights into query formulation patterns [38]. On average, participants in the Search condition formulated queries comprising 4.19 terms. In contrast, participants in the LLM condition issued queries with an average of 6.06 terms. A Chi-Square test revealed a highly significant association between the conditions and the differences in query length patterns ($\chi^2(28) = 61.78, p < 0.001$).

4.1.3 Part-of-Speech Tagging. To explore potential differences in linguistic characteristics, we performed part-of-speech tagging. Figure 4 shows the distribution of tags in the queries across conditions. After adjusting the alpha value using Bonferroni correction, in the LLM condition, several tags, including adverbs (ADV), adpositions (ADP), determiners (DET), auxiliary verbs (AUX), and (VERB) exhibited significantly higher frequencies ($p = 0.051$, $p = 0.046$, $p < 0.001$, $p < 0.001$, $p = 0.019$) than the Search condition. The increased use of adverbs and auxiliary verbs in the LLM queries suggested a more natural language style, potentially influenced by the conversational nature of interactions with LLMs [31]. On the other hand, in the Search condition, usage of proper nouns (PROPN) was significantly higher ($p = 0.034$) than in the LLM condition. This

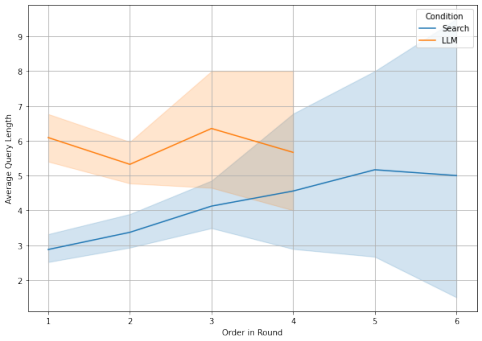


Figure 5: Average number of query terms for successive queries in a round

indicates a greater tendency to perform a keyword-based search using specific entities or locations when interacting with a traditional search engine.

4.1.4 Questions. We explore the categorization of *question* and *non-question* queries. Following Pang and Kumar [30], we defined question queries based on the following criteria:

- Interrogative start: Queries that start with *how, what, which, why, where, when, who, whose.*
- Modal verb start: Queries that start with *do, does, did, can, could, has, have, is, was, are, were, should.* However, an exception is made for queries where the second word is *not.*
- Queries that end with a question mark (?).

Queries not meeting the criteria were classified as *non-question.*

For the Search condition, only 17% were question queries. Conversely, for the LLM condition, 73%, were question queries. A Chi-square analysis yielded statistically significant results between the two types ($\chi^2(1) = 6.37, p = 0.012$). These findings suggest that participants in the LLM condition applied a more conversational style.

4.2 Query Reformulation Strategies

We explore how users reformulate and refine their queries during a given round, focusing on two primary aspects: changes in query length and term repeats, which allows us to understand how participants progressively adapt their queries.

4.2.1 Number of Terms. We investigate how the length of queries changes within round. For each round, we computed the average number of terms in each query in order. Figure 5 shows the average number of terms by query order. Participants in the LLM condition issue an initial query of ∼ 6 terms and maintain this length for subsequent queries. Meanwhile, participants in the Search condition tend to start with shorter (∼ 3) queries and gradually increase. LLM users, favoring longer queries, may indicate a tendency for conversational interactions. Conversely, participants in the Search condition, may reflect an initial focus on keyword-driven retrieval, with subsequent query expansion.

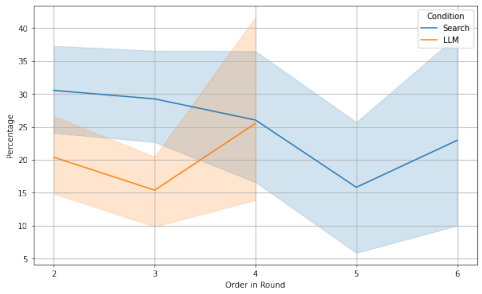


Figure 6: Percentages of term repeats for successive queries in a round

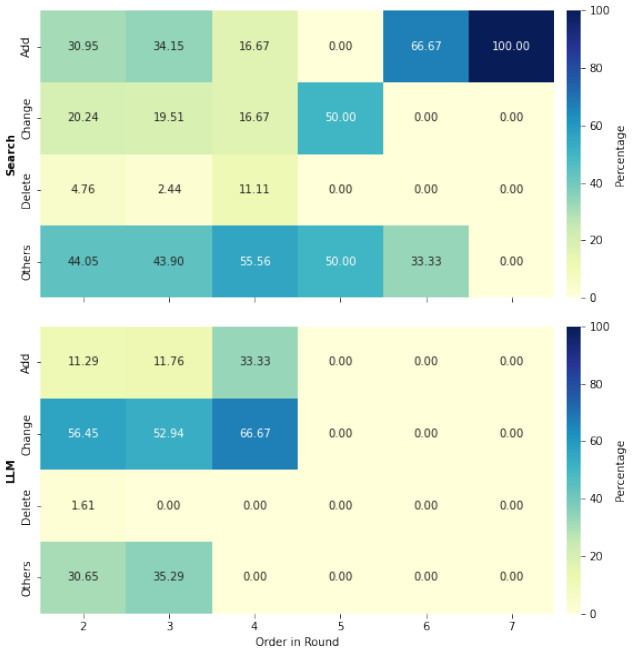


Figure 7: Distribution of syntactic-level types across ordered queries for Search (top), and LLM (bottom) conditions

4.2.2 Term Repeats. We examine term repeats within a round to understand how often users refined their initial queries. We computed the Jaccard similarity percentages [27] of consecutive queries in a round. Figure 6 shows percentages of queries that share identical terms with the previous query in a round. Initially, participants in the LLM condition had generally lower term reuse of around 20%, suggesting a moderate level of query refinement. In the Search condition, participants began with a higher term reuse rate of 30% with a gradual decline as the round progressed, which suggests that participants initially focused on refining their queries, then shifted to queries formulated differently or focused on new clues.

4.3 Query Reformulation Types

Analyzing query reformulation types (QRTs) allows us to infer the user intent in query reformulation. We adopt the QRT taxonomy

Intent Category	Definition	Example Query
Specification (Spec)	Query becomes more specific, narrowing down the search intent	sauf street sign → sauf street sign handicap red x
Generalization (Gen)	Query becomes more general, broadening the search intent	saint james peter adam Hamilton → saint james
Synonym (Syn)	Substitution of a term with its synonym while maintaining the overall meaning	RUE CREVAUX street MAP → RUE CREVAUX street location
Somewhat Relevant (SR)	Intent shifts slightly while remaining somewhat tied to the original query	what language is SAUF in? → how about rue crevaux?
New Topic (New)	Intent shifts significantly to a different subject	what countries have placacentro masisa? → where is 17 de mayo in Chile?
Others (Oth)	Queries that do not fit any category	federal street boston → federal street in (Boston)

Table 2: Intent-level query reformulation types

proposed by Chen et al. [6], which characterizes QRTs at both the syntactic and intent level.

Syntactic changes in consecutive queries, which involve alterations in the structure and composition, are categorized into five types:

- **Add**: New terms are introduced into the query, resulting in an expansion of its content.
- **Delete**: Terms present in the previous query are removed in the current query.
- **Change**: Modifications involve replacing some terms while keeping others unchanged.
- **Repeat**: A query remains identical to the preceding one.
- **Others**: A combination of different changes within a query or the introduction of an entirely new query.

We compute the percent of each syntactic category type for ordered queries in a round. Figure 7 shows the distribution of syntactic-level QRTs for the first, second, etc. query issued in each round. For the Search participants (Figure 7 (top)), we observe that the predominant QRTs were "Add", "Others", and "Change." The high initial rate for "Other" suggests an exploratory intent at the onset. As the task progressed, there was a noticeable rise in the "Add" type, indicating adding details to their queries as they solved the task, suggesting more exploitative behavior.

For the LLM condition (Figure 7 (bottom)), the most frequent QRT was "Change", showing a steady increase as participants progressed. Compared to the Search condition, the "Add" type was less common. This lower occurrence of "Add" suggests that participants in the LLM condition were less inclined to augment their queries with additional terms.

Beyond the syntactic level, we examined the intent level QRT. Rather than measuring *how* users modify queries, the aim is to measure *why* the changes were made to uncover the underlying motivations, evolving information needs, and user goals. Table 2 introduces the six intent categories along with example queries from our dataset.

Two members of the research team performed intent-level categorization. Both researchers individually categorized the queries, then met to reach a consensus on any discrepancies. As shown in (Figure 8 (top)), for participants in the Search condition "Specification", and "New Topic" were the predominant QRTs. Comparing

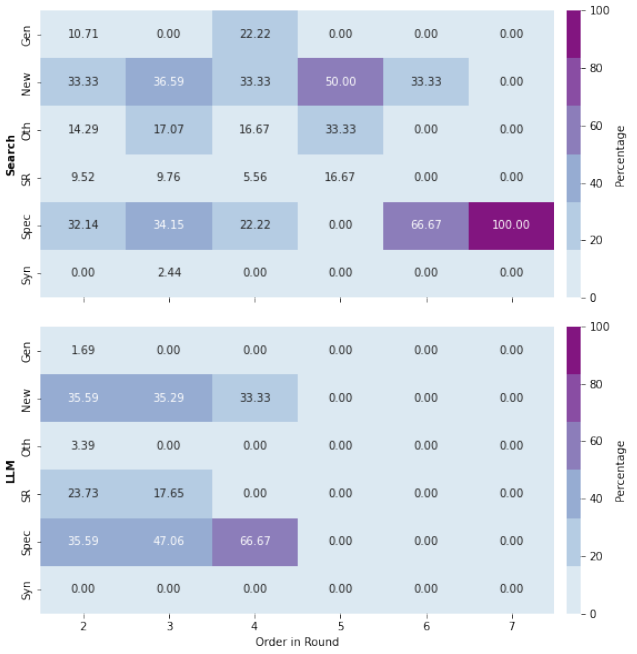


Figure 8: Distribution of intent-level types across ordered queries for Search (top), and LLM (bottom) conditions

these findings with the syntactic changes observed in (Figure 7 (top)), we notice a similar pattern between "Specification" and "Add". This pattern suggests that participants were narrowing down their search intent by adding more details to their query. A parallel trend is observed between "Others" in the syntactic types and "New Topic" as their percentages initially fluctuate but eventually follow a similar pattern.

For the LLM participants, the distribution of intent-level QRTs is shown in (Figure 8 (bottom)). "Specification" was the dominant category. Analysing similar trends with the syntactic changes shown in (Figure 7 (bottom)), there was a steady increase in both 'Specification" and "Add", however "Specification" being much more frequent than "Add." Interestingly, a parallel trend can be observed between "Specification" and "Change." This suggests that while participants using the LLM primarily focused on narrowing down their search intent, they did so without necessarily adding terms to their queries. This behavior differs from the Search condition trend, where both "Specification" and "Add" showed increasing percentages and similar frequencies. This contrast showcases the distinctive user interactions between the two assistants. While Search condition "Specification" often involves query expansion, in the LLM condition, "Specification" primarily shows as query rephrasing.

5 QUALITATIVE FINDINGS

The qualitative findings derive mainly from the responses to the post-study survey and a comparison of the search results returned by each search engine for similar queries.

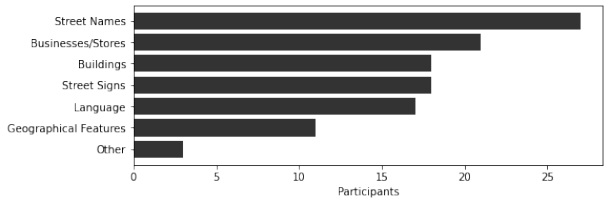


Figure 9: Top clues by participants across conditions

5.1 Open-ended Survey Questions

We conducted a post-study survey with open-ended questions to better understand how participants translated clues into search queries and the challenges they faced.

5.1.1 Clues Identified by Participants. We coded and categorized responses to the question: **What types of clues did you identify that helped you with the task?**. As illustrated in Figure 9, the predominant clue category was street names, followed closely by business names, including store names. Language was also helpful to our participants, particularly to identify non-English speaking countries. Some participants also referred to geographical features like mountains and large bodies of water. The identified clues were consistent across all participants, regardless of the assigned search tool; this aligns with prior work that explored image geolocation [29].

5.1.2 Translating Clues into Search Queries. The post-study survey asked: **How did you translate the clues into search queries?**. Using an inductive, open coding approach [39], we coded and categorized these responses into distinct strategies:

Language Identification. 10 participants from each condition focused on identifying the languages present on signs, buildings, and stores. P13 stated, "I would type in the words I saw, and ask the helper, what language is this in?." Similarly, P18 explained, "I translated some of the clues I saw to English, this way it shows me what the origin of the language."

Locating Street Signs. 10 participants from the Search and 7 from the LLM condition utilized this strategy. They focused on finding street signs in corners and intersections to get closer to the location. P31 explained, "I was searching for street names, trying to identify which neighborhoods the locations were in, for larger cities." P36 also said, "I used the road signs to get a general idea of city and direction of city."

Locating Businesses/Stores. Nine participants from each condition focused on locating businesses, stores, and shops. P37 mentioned, "my strategy was typing company names into the helper."

Describing Geographic Features. A few participants, mostly from the LLM condition, described the geographic features of the location. This included providing details about trees, architecture, and mountains. P13 explained, "I tried describing the environment I was in to the chatbot, but the results were often not good."

Locating Landmarks. In a similar approach to locating businesses, a small number of participants actively searched for large buildings and landmarks. P51 noted, "I searched for a landmark around the area."

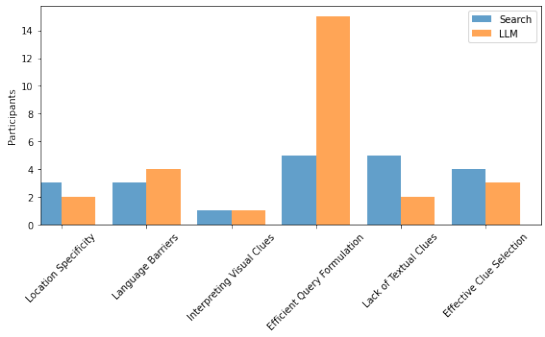


Figure 10: Comparison of challenges faced by our participants in both the Search and LLM conditions

Although these strategies did not differ significantly between the two conditions, analyzing how participants translated visual clues into search queries is important for gaining insights into their approach to image geolocation.

5.1.3 Challenges Identified by Participants. We asked the participants to describe the challenges they faced during the experiment. Only a few participants did not answer or stated they faced no challenges. The responses were coded into in six qualitative classes. Each code, with an example participant response, is given in Table 3.

The distribution of challenges across the experimental conditions is shown in Figure 10. There is notable discrepancy for *Efficient Query Formulation*. While a few participants from the Search condition did mention this challenge, it emerged as the primary obstacle for over half of the participants in the LLM condition. P5 explained, "My biggest challenge was getting the chatbot to understand exactly what I wanted." Similarly, P17 said, "Trying to be concise and precise with my searches using the chatbot was challenging." These participants encountered difficulties in effectively communicating their intent to the LLM-based search engine. Others took a more strategic approach to address this challenge. P46 explained, "I realized I needed to ask less specific questions and go more broadly to get answers." This adaptive strategy reflects participants' attempts to optimize their interactions with the LLM. Participant P55 mentions, "The challenging part was figuring out what I was looking at and translating it to a question that would narrow down answers coming from the chatbot."

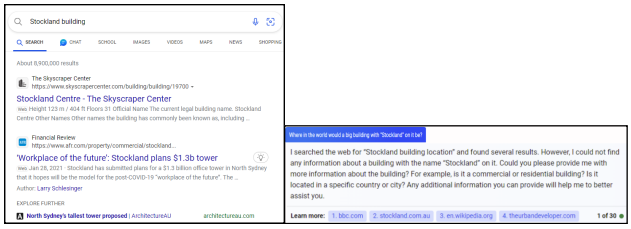
Another challenge worth highlighting was the language barrier. Despite the advanced language capabilities of modern LLMs, the participants had difficulties when formulating language-related questions. P49 stated, "I got a response in Spanish when I typed a Spanish building name but the chatbot didn't answer the question I was asking." Similarly, P28 describes, "I had to rephrase a couple of times, especially If I was asking about phrases in different language than English the chatbot would shoot back a failed search."

5.2 Search Results

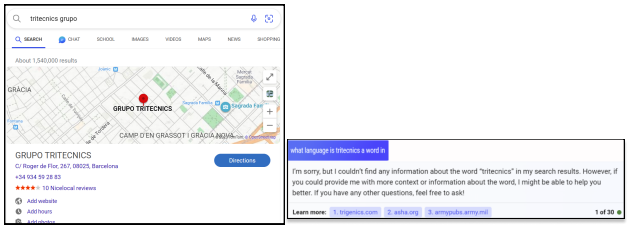
We examined the search queries and the results. For instance, as shown in Figure 11a, when participants searched for "Stockland

Qualitative Code	Description	Example Response
Location Specificity	Precisely specifying and differentiating streets within a city	P1(Search): I did have a hard time. There were many First and Second streets I was on, and it is difficult to distinguish those between the first and second streets of other cities
Language Barriers	Identifying Foreign Words with Non-English alphabets	P13(LLM): My strategy of finding word clues failed if the words I saw were in a language that does not use the English alphabet
Interpreting Visual Clues	Translating visual clues into effective questions or searches	P58(Search): It is hard to try to search the architecture of a building without using an image search
Efficient Query Formulation	Crafting efficient search queries that would yield precise results	P38(LLM): It was challenging to find the correct wording to get the desired result
Lack of Textual Clues	Locations with limited textual clues, for example remote locations	P53(LLM): Sometimes it's very hard to find street names or shop names from the image, especially if the images are from remote locations
Effective Clue Selection	Finding clues that can be described or will generate effective results when searched	P50(LLM): Figuring out what clue to look up, for example, local places were useful, meanwhile large chains are not as useful

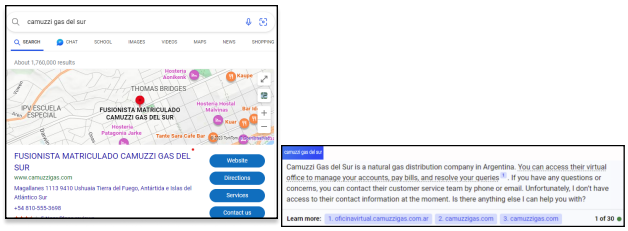
Table 3: Qualitative codes resultant from the coding of challenges described by participants, and example quotes



(a) Round 2 clue: Stockland Building



(b) Round 5 clue: tritecnics



(c) Round 6 clue: Camuzzi gas del sur

Figure 11: Comparison of results obtained from asking about the same clue by participants using different search assistants

building", LLM search did not return any results and requested clarification, while the search engine quickly located the building in Sydney, Australia, in the first few results.

Similar disparities were observed when searching in languages other than English. For instance, when attempting to locate the "tritecnics" company, as shown in Figure 11b, the LLM-based search struggled to translate or make sense of the word, whereas the search engine produced map results for Barcelona, Spain. In Figure 11c, while the LLM-based search successfully replied that "Camuzzi gas del sur" was a gas distribution company in Argentina, the search engine identified the location as the remote city of Ushuaia and presented a map, effectively solving the task for those participants.

6 DISCUSSION

The results demonstrate some key findings: (1) participants using traditional search for assistance outperformed those using the LLM-based search, (2) distinct strategies emerged due to significant variations in the query formulation strategies between the two conditions, and (3) the qualitative findings revealed that participants using the LLM search struggled to effectively formulate their queries.

6.1 Differences in Performance

In response to **(RQ1)**, the results showed that participants using the traditional search outperformed those using the LLM search. We believe that a substantial portion of the performance difference can be explained by the difference in results for similar queries, as shown in Section 5.2. Specifically, as shown in Figure 11b and Figure 11c, the search engine provided a map, effectively directing our participants to the exact location. LLM-based search should not only provide creative answers, but the same features as a traditional search engine, including maps. As evident from the examples in our study, LLM search did not always return results that should have been within its capabilities. Perhaps, the participants needed to explicitly request maps or specific directions to effectively utilize these features; however this was not necessary for traditional search. This highlights the challenge of crafting effective prompts,

Query	LLM Response	Score
countries in southern hemisphere with mountains	The Southern Hemisphere is home to many mountain ranges, including the Andes, the Sierra Nevada de Santa Marta, and the Cordillera Oriental and Cordillera Occidental in South America. The highest peak in the Americas, Aconcagua, is also located in the Andes. In Africa, Mount Kilimanjaro is the tallest mountain and one of the world's largest volcanoes. The Rwenzori range of mountains in Africa were created during the formation of the Great Rift Valley. New Zealand's South Island has more than 24 mountains with elevations of 9,000 feet or more[1]. I hope this helps!	605
patas para arriba location	(Original Spanish) La expresión "patas para arriba" se utiliza para describir una situación en la que todo está desordenado o confuso. ¿Hay algo más en lo que pueda ayudarte? (Translated) The expression "upside down" is used to describe a situation in which everything is disordered or confusing. Is there anything else I can help you with?	1240
24seventaxi location	I'm sorry but I couldn't find any information about 24seventaxi location. Could you please provide me with more details about what you are looking for?	1773
I need to know where the street H. de Magallanes is located in the world	I'm sorry but I couldn't find any information about H. de Magallanes street. Could you please provide more details about the location or any other information that might help me find it?	18

Table 4: Examples where participants developed inaccurate mental models, resulting in the formulation of single, poor-quality queries, which led to low performance

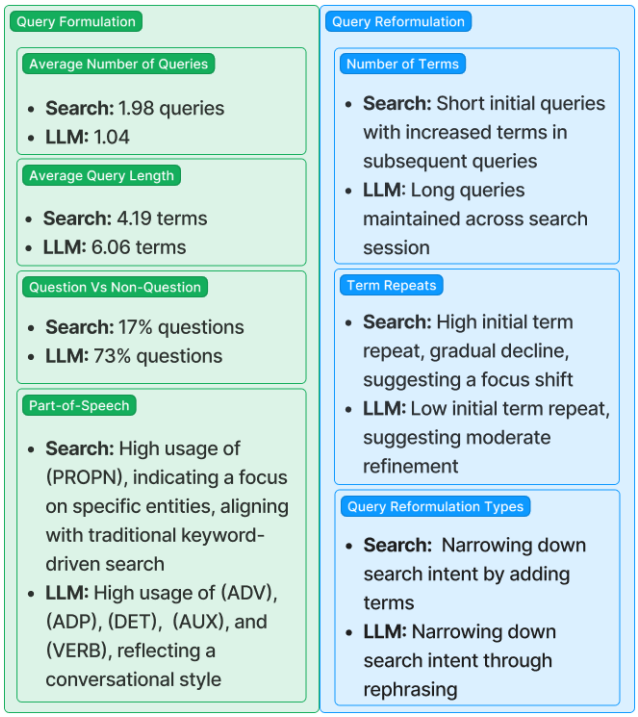


Figure 12: Summary findings of query formulation and reformulation analysis.

which, as research has shown [47], significantly influences the output of LLM-based methods. While "Learn more" links were included with each answer, we found that none of the participants took advantage of this feature. This raises questions about the perceived affordances of LLMs compared to traditional search engines, as the integration of similar features in LLMs may not be as intuitive, as our study suggests. This underutilization of help links in LLM responses highlights the challenges of transitioning from traditional search engines to LLM-based search and aligns with Gibson's theory [10] emphasizing the interaction between users and their technological environment, which shapes the possibilities and constraints for action.

In response to **(RQ3)**, our qualitative analysis provided valuable insights into the performance disparities. Over half of the participants using the LLM-based search expressed difficulties in formulating their queries. Some expressed hesitation, while others found it challenging to form queries that effectively communicated their information needs. Participants also struggled to formulate queries in different languages. Despite stated support of LLM search for many languages, there were reported instances of failed results. This emphasizes the challenge of effectively prompting the LLM to comprehend and respond to queries in diverse languages. These challenges were less prevalent when using the search engine. These observations align with our query formulation analysis. In Section 4.1.1, we noted that the average number of queries for LLM participants consisted of a single query. Figure 3 showed that LLM participants were less inclined to reformulate or issue more queries throughout the task compared to Search participants.

While it has been shown that individuals can quickly build mental models when interacting with LLM chatbots [16], the quality of these models remains uncertain, especially considering the relatively new nature of LLM technology. Participants may not have had the experience to develop accurate mental models of LLM capabilities. In the absence of well-defined mental models, users struggle to predict outcomes or make sense of their interactions with LLMs, leading to instances where users pose vague and poorly defined queries while expecting the LLM to respond appropriately [37, 47]. Table 4 demonstrates a few examples of this behavior. It is plausible that these mental models influenced the participants to attempt only a single query, potentially resulting in worse performance.

6.2 Differences in Query Formulation & Reformulation Strategies

Figure 12 presents a summary of our findings into query formulation and reformulation strategies, which is directly related to **(RQ2)**. As described in Section 4.1, our study showed significant differences across all four formulation metrics. Participants in the Search condition issued shorter queries and a significantly increased use of proper nouns including places, store names, and streets. These findings align with prior research [18, 24], indicating that individuals are accustomed to keyword-based search from a lifetime of experience using traditional search engines. In contrast, participants using the LLM-based search were inclined to issue longer and more natural language queries. These findings suggest that they adhered to the perceived norms of a conversational user interface [8].

Several research efforts have explored the topic of query reformulation for traditional web search [2, 13, 20]. In our study, we adopted an existing taxonomy for categorizing both the syntactic and intent-based types of query reformulation, detailed in Section 4.2. Our investigation revealed a notable trend. Although "Specification" or

the act of narrowing down the scope and intent of a search, was the dominant category across both conditions, it aligned with different syntactic categories. In the Search condition, participants often expanded their queries by adding terms, a common behavior observed in prior research [6, 7]. However, participants using the LLM search frequently narrowed down their search intents by paraphrasing their initial queries. This observation prompts a critical question of whether the development of new query reformulation taxonomies specific to LLMs could provide a better framework for understanding and characterizing user behavior.

6.3 Geolocation Sensemaking Strategies

As described in Section 5.1.2, participants employed diverse strategies when searching. In geolocation tasks, participants make sense of the clues within the images by using *internal* or *external* knowledge representations [29]. Our focus was on external knowledge representations cultivated through their searching. Therefore, we did not ask about internal knowledge, such as participants' cultural backgrounds or travel history.

In geolocation tasks, participants engage in sensemaking to interpret visual cues within images. Participants demonstrate adaptability in their approach, drawing from Pirolli and Card's sensemaking model [33], which encompasses both top-down and bottom-up approaches. The top-down approach involves initiating the process with a theory or a broader concept and then seeking data to substantiate it. In this context, an example of the top-down approach is the strategy of language identification. Participants effectively employ this strategy by querying the assistant about the language's origin, which narrows the search scope to specific global regions. Similarly, some participants employed sensory sensemaking as a top-down approach by describing sensory aspects of the location to the assistant. However, the limited effectiveness of this strategy suggests that relying solely on sensory cues may not be sufficient for precise geolocation. The bottom-up approach, on the other hand, entails gathering data first and progressively forming a theory based on the available information. Examples of this approach included using street signs, buildings, and landmarks as reference points, facilitating the identification of cities then neighborhoods. The strategies identified in this study provide insights into user behavior and align with the cognitive processes driving geolocation sensemaking described by prior work [41]. Understanding how participants construct mental models based on image clues and apply sensemaking processes, enhances our ability to provide effective support and guidance in geolocation tasks.

6.4 Limitations

Amongst the study limitations was the latency in LLM-based search. Although the latency was only a few seconds, it could have disrupted the conversational flow, potentially affecting satisfaction and engagement during the task. Our study included participants with diverse backgrounds, education levels, ages, and degrees of technical literacy. This inherent diversity may have influenced how participants interacted with both the image geolocation task and the search engines. Furthermore, it's essential to acknowledge that image geolocation tasks have historically been conducted by expert image analysts. In our study, we did not explicitly categorize participants based on their levels of expertise in geolocation. Lastly, our evaluation did not include the assessment of specific metrics such as search engine result pages (SERPs), clicks, or other performance indicators that could offer a more comprehensive view of the effectiveness of LLM-based search in image geolocation tasks.

7 CONCLUSION AND FUTURE WORK

This study offered valuable insights into differences in strategies and user behaviors when using traditional compared to LLM-based search for image geolocation. We examined the differences in performance, query formulation, and the sensemaking strategies employed by participants in these two conditions. Despite the growing capabilities of LLMs, the results reveal that participants using traditional search engines outperformed those relying on LLM-based search. An in-depth exploration of the distinct query formulation strategies utilized by participants mostly explained the performance difference, as evidenced by our qualitative findings, with query formulation being identified as the most challenging aspect of the experiment. Additionally, we observed a tendency among participants using LLMs to engage in fewer multi-query search sessions, possibly reflecting uncertainties surrounding LLM capabilities and the perceived affordances associated with LLM interface.

Our findings can extend beyond the geolocation domain, providing initial insights into user interactions with LLMs in real-world applications and prompting more research on human-centered design of LLM interfaces, with a focus on understanding how users form mental models of LLMs. To achieve more useful LLM interfaces, it is necessary to first develop a better understanding of query formulation strategies and behavior. Extensive prior research about traditional search provides a solid foundation for exploring query formulation strategies. Our work presented in this paper begins to extend this research based on the novel capabilities and the conversational nature of LLM-based search; however, more research in this area is needed. The second component of making LLM interfaces more usable is to teach novices how to effectively prompt. Emerging systems like AI Chains [46], MemorySandbox [14], Feedback Buffet [28], and PromptMaker [19] are at the forefront, making LLMs more comprehensible and user-friendly through the use of templates [19, 28] and procedural guidance [14, 15, 46]. These tools are designed to assist novice users in prompt creation by integrating visual problem representation, incorporating partial prompts, and providing user friendly interfaces that facilitate easy iteration based on the LLM output. These advancements represent a leap towards a future in which user interactions with language models become more intuitive, efficient, and user-friendly.

ACKNOWLEDGMENTS

Thanks to Andrea Brandt for assisting with the user study. This research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Leonard Adolphs, Shehzaad Dhuliawala, and Thomas Hofmann. 2021. How to Query Language Models? arXiv:2108.01928 [cs.CL]

[2] Anne Aula. 2003. Query Formulation in Web Information Search.. In *ICWI*. International Conference WWW/Internet, Algarve, Portugal, 403–410.

[3] Cory Barr, Rosie Jones, and Moira Regelson. 2008. The Linguistic Structure of English Web-Search Queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii) *(EMNLP '08)*. Association for Computational Linguistics, USA, 1021–1030.

[4] Simon Bilel Jegham, dim fort. 2021. *GeoGuess*. MIT Licensed. https://github.com/GeoGuess/GeoGuess Accessed: 2023.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[6] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a Better Understanding of Query Reformulation Behavior in Web Search. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 743–755. https://doi.org/10.1145/3442381.3450127

[7] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating query reformulation behavior of search users. In *China Conference on Information Retrieval*. Springer, China, 39–51.

[8] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, USA, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[10] James J Gibson. 2014. *The ecological approach to visual perception: classic edition*. Psychology press, Online.

[11] Jutta Haider and Olof Sundin. 2019. *Invisible search and online search engines: The ubiquity of search in everyday life*. Taylor & Francis, London.

[12] James Hays and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*. IEEE, CVPR, Alaska, USA, 1–8.

[13] Jeff Huang and Efthimis N. Efthimiadis. 2009. Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 77–86. https://doi.org/10.1145/1645953.1645966

[14] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 97, 3 pages. https://doi.org/10.1145/3586182.3615796

[15] Ziheng Huang, Kexin Quan, Joel Chan, and Stephen MacNeil. 2023. CausalMapper: Challenging Designers to Think in Systems with Causal Maps and Large Language Model. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) *(C&C '23)*. Association for Computing Machinery, New York, NY, USA, 325–329. https://doi.org/10.1145/3591196.3596818

[16] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: investigating social facilitation in human-machine team creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama,Japan, 1–16.

[17] Rahul J Jadhav, Om Prakash Gupta, and Usharani T Pawar. 2011. Significant role of search engine in higher education. *International Journal of Scientific & Engineering Research* 2, 4 (2011), 1–5.

[18] Bernard J Jansen and Amanda Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information processing & management* 42, 1 (2006), 248–263.

[19] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-Based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. https://doi.org/10.1145/3491101.3503564

[20] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning User Reformulation Behavior for Query Auto-Completion. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 445–454. https://doi.org/10.1145/2600428.2609614

[21] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? arXiv:1911.12543 [cs.CL]

[22] Seungun Kim, Masaki Matsubara, and Atsuyuki Morishima. 2022. Image Geolocation by Non-Expert Crowd Workers with an Expert Strategy. In *2022 IEEE International Conference on Big Data*. IEEE xplore, Osaka, Japan, 4009–4013. https://doi.org/10.1109/BigData55660.2022.10020932

[23] Rachel Kohler, John Purviance, and Kurt Luther. 2017. Supporting Image Geolocation with Diagramming and Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (Sep. 2017), 98–107. https://doi.org/10.1609/hcomp.v5i1.13296

[24] Dirk Lewandowski. 2008. Search engine user behaviour: How can users be guided to quality content? *Information Services & Use* 28, 3-4 (2008), 261–268.

[25] Xinyi Li, Bob J.A. Schijvenaars, and Maarten de Rijke. 2017. Investigating queries and search failures in academic search. *Information Processing & Management* 53, 3 (2017), 666–683. https://doi.org/10.1016/j.ipm.2017.01.005

[26] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. 2015. Learning deep representations for ground-to-aerial geolocalization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE xplore, Boston, MA, USA, 5007–5015. https://doi.org/10.1109/CVPR.2015.7299135

[27] Chang Liu, Xiangmin Zhang, and Wei Huang. 2016. The exploration of objective task difficulty and domain knowledge effects on users' query formulation. *Proceedings of the Association for Information Science and Technology* 53 (12 2016), 1–9. https://doi.org/10.1002/pra2.2016.14505301063

[28] Stephen MacNeil, Andrew Tran, Joanne Kim, Ziheng Huang, Seth Bernstein, and Dan Mogil. 2023. Prompt Middleware: Mapping Prompts for Large Language Models to UI Affordances. arXiv:2307.01142 [cs.HC]

[29] Sneha Mehta, Chris North, and Kurt Luther. 2016. An exploratory study of human performance in image geolocation tasks. In *GroupSight Workshop on Human Computation for Image and Video Analysis*, Vol. 308. HCOMP 2016, Austin, TX (USA), 3–4.

[30] Bo Pang and Ravi Kumar. 2011. Search in the lost sense of "query": Question formulation in Web search queries and its temporal changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, USA, 135–140.

[31] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Alfred Sliwa, Ahmet Aker, and Norbert Fuhr. 2021. Starting Conversations with Search Engines - Interfaces That Elicit Natural Language Queries. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 261–265. https://doi.org/10.1145/3406522.3446035

[32] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How Context Affects Language Models' Factual Predictions. arXiv:2005.04611 [cs.CL]

[33] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[34] Amudha Ravi Shankar, Jose Fernandez-Marquez, Gabriele Scalia, Maria Rosa Mondardini, and Giovanna Di Marzo Serugendo. 2019. CROWD4EMS: A CROWDSOURCING PLATFORM FOR GATHERING AND GEOLOCATING SOCIAL MEDIA CONTENT IN DISASTER RESPONSE. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-3/W8 (08 2019), 331–340. https://doi.org/10.5194/isprs-archives-XLII-3-W8-331-2019

[35] 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* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[36] S Salehi, J Tina-Du, and H Ashman. 2018. Use of Web search engines and personalisation in information searching for educational purposes. iRInformation Research.

[37] Hariharan Subramonyam, Christopher Lawrence Pondoc, Colleen Seifert, Maneesh Agrawala, and Roy Pea. 2023. Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces. *arXiv preprint arXiv:2309.14459* none, none (2023), 10 pages.

[38] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. 2011. TwitterSearch: A Comparison of Microblog Search and Web Search. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China)

(WSDM '11). Association for Computing Machinery, New York, NY, USA, 35–44. https://doi.org/10.1145/1935826.1935842

[39] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[40] Paul Thomas, Bodo Billerbeck, Nick Craswell, and Ryen W White. 2019. Investigating searchers' mental models to inform search explanations. *ACM Transactions on Information Systems (TOIS)* 38, 1 (2019), 1–25.

[41] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting expert image geolocation with crowdsourcing and shared representations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[42] Yiwei Wang, Jiqun Liu, Soumik Mandal, and Chirag Shah. 2017. Search successes and failures in query segments and search tasks: A field study. *Proceedings of the Association for Information Science and Technology* 54, 1 (2017), 436–445.

[43] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL]

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., New Orleans, USA, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[45] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *Computer Vision – ECCV 2016*. Springer International Publishing, Amsterdam, The Netherlands, 37–55. https://doi.org/10.1007/978-3-319-46484-8_3

[46] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[47] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[48] Guido Zuccon, Bevan Koopman, and Joao Palotti. 2015. Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, March 29-April 2, 2015. Proceedings 37*. Springer, Vienna, Austria, 562–567.