

From Web Search to AI Search: A Comparative Analysis of Information-Seeking Behaviour in Traditional Search Engines and Conversational Systems Based on Large Language Models

Arianna Di Vittorio

Department of Economics, Management and Business Law,
University of Bari “Aldo Moro”, Bari, Italy

Di Gioia Vito Alessandro

MatiPay Srl, Bari, Italy

Approved: 08 October 2026

Posted: 10 October 2026

Copyright 2026 Author(s)

Under Creative Commons CC-BY 4.0

OPEN ACCESS

Cite As:

Di Vittorio, A., & Di Gioia, V.A. (2026). *From Web Search to AI Search: A Comparative Analysis of Information-Seeking Behaviour in Traditional Search Engines and Conversational Systems Based on Large Language Models*. ESI Preprints.

<https://doi.org/10.19044/esipreprint.10.2026.p187>

Abstract

AI search is changing how users seek information: instead of typing keywords and clicking through links, users increasingly ask complex questions and obtain direct answers on the results page. This shift, together with the rise of “zero-click search” and of chatbots based on Large Language Models (LLMs), is reshaping the customer experience of information seeking. This paper examines how LLM-based search systems are changing the way users retrieve information online, moving the dominant paradigm from keyword queries and ranked lists of links to direct answers in natural language. The study offers a structured, non-systematic narrative review comparing traditional search engines (represented by Google Search) and conversational AI tools (represented by ChatGPT and GPT-3/3.5-based search tools). It synthesises eight independent studies published between 2023 and 2026: three randomised controlled experiments, two large-scale observational analyses of real-world behaviour and three studies measuring content reliability, complemented by an industry benchmark (Vectara Hallucination Leaderboard). The analysis focuses on five key performance indicators (KPIs): number of queries per task (KPI 1), task success

rate/accuracy (KPI 2), total time to task completion (KPI 3), rate of consultation of external sources/outbound clicks as a proxy for cross-verification (KPI 4), and the hallucination rate of AI systems compared with the rate of unreliable results in traditional search (KPI 5). AI systems consistently reduce task completion time, whereas the reduction in the number of queries depends on the task. Accuracy improves only when the generated output is correct; when it is wrong, accuracy drops sharply, consistent with a markedly lower consultation of external sources. The risk of error exists in both paradigms under specific conditions of use. The five KPIs should therefore be read as an interdependent system, with implications for the design of hybrid search tools in human-computer interaction and information retrieval.

Keywords: AI search; information-seeking behaviour; Large Language Models; search engines; human-computer interaction; key performance indicators; New Customer Experience

1. Introduction

1.1. The changing information search phase of purchasing behaviour: from keywords to technical conversation

Information seeking, until recently conducted almost exclusively through web search engines, has in recent years been increasingly complemented by conversational artificial intelligence platforms, with a direct and observable impact on how the information supporting commercial decisions is retrieved, synthesised and verified. Since the public release of ChatGPT in November 2022, the adoption of conversational AI as an information channel has grown substantially, alongside and partly overlapping with traditional search engines.

The paradigm underlying Google Search - indexing and lexical matching of the query, returning a ranked set of hyperlinks that users must explore on their own - differs structurally from the conversational paradigm of systems based on Large Language Models (LLMs), in which a request formulated in natural language receives a synthetic, discursive answer.

Before the advent of AI, users retrieved information by typing keywords into the search bar and clicking through links. Today they type a request and obtain elaborated, contextualised answers directly on the results page, often followed by further questions and extended reasoning cycles (deep-research agents). This produces a first major change, known as “zero-click search” (Table 1): a search that ends on the results page without the user clicking on any result. A new buyer journey thus emerges, shifting from keyword search to technical conversation.

Table 1: Impact of “zero-click search”

Search metric	Current average value	Impact and context
Google "zero-click" searches (USA)	~58.5%	More than half of searches end without the user clicking on any website.
Google "zero-click" searches (Europe)	~59.7%	Trend mirroring the US, driven by direct answers.
Click-through rate when AI Overviews are shown	Only 8%	When Google displays the AI summary at the top, users click a traditional link in only 8% of visits (vs. 15% without a summary).
Google AI Mode sessions ending without a click	93%	In Google AI Mode only 6-8% of sessions lead the user to an external website (estimate 92-94%).

Source: Fishkin (2024), SparkToro/Datos, 2024 Zero-Click Search Study (US/EU data); Chapekis & Lieb (2025), Pew Research Center (AI Overviews data); Harsel (2025), Semrush (Google AI Mode data).

Zero-click searches can still benefit firms that invest in search visibility in four situations: the user reformulates the query, refining the search intent; the user looks for short, concise information and is satisfied with the SERP (Search Engine Results Page), but notes the website from which the information is taken; the user contacts the firm directly; or the user moves on to a dedicated app, such as Google Maps or Wikipedia. Zero-click search therefore serves needs other than in-depth exploration, and by optimising for this search intent search engines attract customers more directly and personally, answering faster and more pertinently and turning from “search” engines into “answer” engines.

This transition was gradual: in 2022 and 2023 traditional search engines retained an exclusive role, whereas with the introduction of Google's Search Generative Experience at I/O 2023, which evolved into AI Overviews in May 2024, hybrid interaction combining keyword search and generative synthesis began to spread on a mass scale (Table 2).

Table 2: Timeline from keyword search to conversational search (2022-2026)

Year	Source	Reported data / information
2022	Datos / SparkToro	DuckDuckGo recorded 106 million searches per day in a still keyword-based market; ChatGPT was launched in late November 2022, without web search capability.
2023	Google I/O 2023	Launch of Search Generative Experience; in a sample of 100 013 keywords, 18 455 (18.45%) already triggered AI snippets; almost 60% of Google searches were already long-tail queries of 4+ words.
2024	Google I/O 2024	Google searches grew by 21.64% in a single year, confirming that AI-assisted search complements rather than replaces traditional search.
	SE Ranking, tracking AI	Growth from 18.55% of SERPs in Q3 2024 to 49.92% by December 2024.

2025	Overviews	
	BrightEdge	Total search impressions grew by about 49% since the launch of AI Overviews (May 2024), while click-through fell by almost 30%.
	Nectiv Digital (analysis of over 8 500 real prompts)	ChatGPT's internal search queries average 5.48 words, 61% longer than the Google average of 3.4 words (Semrush data).
	Search Engine Land	Confirms an average length of 5.48 words, with 77% of queries of 5 words or more.
	Semrush	Average Google query $\approx$ 4 words; in AI Mode it lengthens to 7 words; with ChatGPT as a conversational collaborator the average rises to 23 words.
	Datos / SparkToro	Google holds 93.57% of the search market ($\sim$ 14 billion queries/day); ChatGPT processes about 37.5 million "search-like" prompts per day, 0.25% of the market.
	iPullRank	AI search queries (query fan-out) average 70-80 words, versus 3-4 in traditional searches.
	Similarweb	ChatGPT: 5.8 billion monthly visits, 489 million unique users, fifth most visited website worldwide, 115.9% growth by the end of 2024; however, 95% of ChatGPT users still also use Google, and only 23 million use ChatGPT exclusively.

Source: industry surveys and reports listed in Table 2. None of the sources in the figure is academic: the Google I/O rows report company announcements and the others industry (SEO/marketing) reports, useful only as orders of magnitude. These sources are not among the studies included in the review (Section 2.1.1).

The academic literature has begun to document this behavioural shift empirically. A first strand of experimental studies has compared users' preferences and performance with search engines and conversational LLMs, finding that direct questions are still entrusted to traditional search engines, whereas tasks requiring understanding are delegated to conversational systems (Caramancion, 2024). A second strand has confirmed that the emergence of generative chatbots represents a significant change in information retrieval, observable across users of different ages and educational levels.

AI makes many links irrelevant because it immediately shows users what they can or should do differently: value is shifting from "optimised" content to content that "provides answers". Searches for information on complex products or services are therefore bound to involve AI platforms as well, since these are increasingly part of everyday life, generating purchase journeys that are better informed and richer in content (Table 3).

Table 3: Evolution of information search by buyers (2024-2026)

Year	Source	Data relevant to B2B
2024	Forrester, Buyers' Journey Survey	89% of B2B buyers have adopted generative AI and cite it as one of the main sources of self-directed information at every stage of the purchasing process.
2025	6sense, global study (~4 000 B2B buyers), G2, Buyer Behaviour Report	94% of buyers now use LLMs during the purchase journey, compared with 68% who two years earlier stated that generative AI had no impact on their purchasing process. GenAI chatbots are becoming the most influential source for building B2B vendor shortlists (17.1%), ahead of software review sites (15.1%), vendor websites (12.8%) and peer recommendations (8.9%).
2026	Omniscient / Wynter, "From Prompt to Purchase"	The purchase journey is now a loop: buyers search on Google, compare on an LLM, and buy from the vendor's website or by contacting the vendor.
	Forrester, Buyers' Journey Survey (18 000 global buyers)	B2B buyers now use AI tools to research products (54%), compare vendors (55%) and build internal business cases before contacting a vendor (47%).
	TrustRadius	72% of buyers now encounter Google AI Overviews as part of their research, and 90% of them click to verify the sources cited in those AI answers: human verification remains central even when AI generates the first answer.

Source: B2B/fintech industry surveys and reports

2. Structured narrative review and quantitative formalisation of search-behaviour Key Performance Indicators

Building on the contrast between the two paradigms outlined in Section 1, the paper addresses five research questions, one for each KPI. RQ1: do users need fewer interactions to achieve a search goal with an AI system (KPI 1)? RQ2: does greater efficiency translate into greater accuracy, or does it introduce risks related to overreliance (KPI 2)? RQ3: to what extent is efficiency reflected in task completion time (KPI 3)? RQ4: how does a conversational system affect the independent consultation of external sources (KPI 4)? RQ5: what is the order of magnitude of the risk of receiving incorrect information in the two paradigms, and what factors does it depend on (KPI 5)?

2.1. Research methodology

2.1.1. Study selection criteria

The paper adopts a structured narrative synthesis, reported following, as far as applicable, the SWiM (Synthesis Without Meta-analysis) guideline (Campbell et al., 2020): the heterogeneity of tasks, measures and populations

precludes formal meta-analytic pooling (Borenstein et al., 2009). The review analyses eight independent studies published between 2023 and 2026, organised in two strands. Strand A (search behaviour) includes studies that: a) measure behaviour objectively, through system logs or clickstream data; b) explicitly compare a traditional search system with an LLM-based conversational system; c) provide sufficient quantitative data to compute an effect measure.

Strand B (content reliability) includes studies that: a) verify, through a documented procedure, the accuracy of outputs or the reliability of the sources returned; b) report the error rate with its denominator. Strand A comprises three randomised experiments (Xu et al., 2023; Spatharioti et al., 2025; Kaiser et al., 2025a, 2025b) and two large-scale observational studies (Iannelli & Ai, 2026; Shi et al., 2026); strand B comprises Chelli et al. (2024), Dahl et al. (2024) and Aslett et al. (2024). The Vectara Hallucination Leaderboard is used as a contextual benchmark, not as an included study.

Studies were identified through purposive selection rather than a systematic protocol (e.g., PRISMA), searching Google Scholar, arXiv and the ACM Digital Library with combinations of the keywords («ChatGPT» OR «large language model» OR «generative AI» OR «AI search») AND («search engine» OR «web search» OR «Google») AND («user behavior» OR «experiment» OR «clickstream» OR «hallucination»), supplemented by citation searching; the last search was run on 3 October 2026. The number of records screened was not recorded: the review does not claim to be exhaustive and selection bias cannot be ruled out.

Figs. 1, 2 and 3 rely on industry sources (Fishkin, 2024; Chapekis & Lieb, 2025; Harsel, 2025; Google I/O 2023 and 2024; SE Ranking, BrightEdge, Nectiv Digital, Search Engine Land, Datos/SparkToro and iPullRank) that are not peer-reviewed and are continuously updated. They are used only for context, indicate the order of magnitude of the phenomenon and do not enter the KPI synthesis, which is based solely on the studies in strands A and B. The values of the Vectara Hallucination Leaderboard were checked on 5 October 2026.

The bibliographic verification of non-peer-reviewed sources, carried out on 5 October 2026, confirmed: Iannelli and Ai (2026), arXiv:2607.04282, v1 of 5 July 2026 and v3 (current version) of 25 August 2026, authors affiliated with Scrunch AI, New York (USA); Shi, Zhu and Gu (2026), arXiv:2607.07652, v1 of 8 July 2026, authors affiliated with Bocconi University, Milan (Italy), also deposited on SSRN (abstract no. 7035298, 22 July 2026); Xu, Feng and Chen (2023), arXiv:2307.01135. The publication status of each study is summarised in the methodological note following the References.

2.1.2. Included studies

Xu et al. (2023) compare ChatGPT (GPT-3.5-turbo) and Google Search in an online between-subjects experiment with 95 participants recruited via Prolific (48 in the ChatGPT group, 47 in the Google Search group), using three tasks of increasing complexity: a factual question, compiling a list of five websites for booking a flight, and fact-checking three statements about the 2009 UN Climate Change Conference in Copenhagen. Spatharioti et al. (2025) run two experiments on Amazon Mechanical Turk (90 and 120 participants) in which participants identify which of two SUVs has the better ratio of cargo space to overall length, using a Bing-based tool or a tool based on GPT-3.5 (first experiment) or GPT-3 (second experiment, which also tests visual low-confidence cues). Kaiser et al. (2025a, 2025b), for the Nuremberg Institute for Market Decisions, randomly assign 1 526 US participants to ChatGPT or Google for a consumer decision (smartphones or theme parks), tracking searches, time and websites visited. Iannelli and Ai (2026), whose authors are affiliated with Scrunch AI, use an opt-in panel linking users' conversations with AI assistants to their searches and web browsing. Shi et al. (2026) analyse desktop clickstream data from 45 386 US households (Comscore panel, October 2024 - July 2025) and estimate, with a difference-in-differences design, the intention-to-treat effect of three staggered expansions of access to ChatGPT Search.

In strand B, Chelli et al. (2024) and Dahl et al. (2024) measure the hallucination rate of several LLMs in open-ended factual retrieval, in the medical-bibliographic and legal domains respectively; Aslett et al. (2024, Nature) analyse, in five experiments with over 14 000 participants, the reliability of Google Search results as a function of query formulation.

Not all studies meet the criteria in full. Kaiser et al. (2025a, 2025b) report the size of the four experimental cells (ChatGPT-smartphone $n=412$; ChatGPT-theme parks $n=382$; Google-smartphone $n=359$; Google-theme parks $n=373$; total $N=1\,526$), but not all statistical tests: the secondary analyses reported here use these cell sizes and counts reconstructed from rounded percentages and are therefore approximate. The outbound-click comparison of Shi et al. (2026) is descriptive; Iannelli and Ai (2026) do not report the panel size; Aslett et al. (2024) measure the reliability of returned sources through the domain's NewsGuard score.

2.2. KPI formalisation and effect measures

This section formalises the five KPIs and the effect measures adopted, using a unified notation: the subscript “s” denotes the system (AI or traditional search), “ N_s ” the number of participants assigned to that system (search occasions in clickstream data are denoted by O_s), the indices “i” and “k” the participant and the task respectively, and the index “r” the position of

a result on the results page (Equation 5). Within the equations, the label “trad” denotes traditional search.

KPI 1 - Mean number of queries per task. Equation 1 defines the sample mean, the difference between systems and the regression model, where q_i is the number of queries of participant i (q_{ik} when task k is distinguished). As this is a count, a Poisson generalised linear model with log link is adopted, with a fixed task effect (γ_k) and a random participant intercept (u_i); AI_i equals 1 if participant i is assigned to the AI system and 0 otherwise, and $\exp(\beta_1)$ is the ratio of the query rates of the two systems (rate ratio). In the presence of overdispersion, negative binomial regression is used. The model follows the specification of Spatharioti et al. (2025); since the review works on aggregate data, it is not estimated by the authors and is proposed as a reference for future replications on individual-level data.

Equation 1 - Formalisation of KPI 1 (mean number of queries per task, difference between systems, mixed-effects Poisson model)

$$\begin{aligned}\bar{Q}_s &= \frac{1}{N_s} \sum_{i=1}^{N_s} q_i \\ \Delta Q &= \bar{Q}_{AI} - \bar{Q}_{trad} \\ q_{ik} &\sim \text{Poisson}(\mu_{ik}) \\ \log \mu_{ik} &= \beta_0 + \beta_1 AI_i + \gamma_k + u_i \\ u_i &\sim N(0, \sigma_u^2)\end{aligned}$$

KPI 2 - Task success rate (accuracy). The indicator is defined in binary form, as the proportion of responses $\hat{y}_{ik}$ matching the reference answer y_{ik} , where K_i is the number of tasks or decisions assessed for participant i , and in continuous form, as the mean score on a 0-10 scale (Xu et al.), reported in the figures as a percentage ($10 \times \bar{A}$) (Equation 2).

Equation 2 - Formalisation of KPI 2 (task success rate, binary and continuous form)

$$\begin{aligned}A_s &= \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{1}{K_i} \sum_{k=1}^{K_i} \mathbb{1}(\hat{y}_{ik} = y_{ik}) \\ \bar{A}_s &= \frac{1}{N_s} \sum_{i=1}^{N_s} a_i, \quad a_i \in [0, 10]\end{aligned}$$

KPI 3 - Total time to task completion. In addition to the mean time per system, Equation 3 defines efficiency per query as the ratio of mean time to the mean number of queries of the system, which can be computed from published aggregate data.

Equation 3 - Formalisation of KPI 3 (total time to completion and efficiency per query)

$$\bar{T}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} t_i$$

$$E_s = \frac{\bar{T}_s}{\bar{Q}_s}$$

KPI 4 - Rate of consultation of external sources (proxy for cross-verification). The indicator is computed as the proportion of participants consulting at least one external source (V , experimental data; N_s participants) and as the proportion of search occasions with at least one outbound click (C , clickstream data); since an occasion corresponds to a session for ChatGPT and to a query for Google, the two forms are not comparable (Equation 4).

Equation 4 - Formalisation of KPI 4 (share of users consulting external sources and share of search occasions with outbound clicks)

$$V_s^{\text{users}} = \frac{n_s^{\text{external}}}{N_s} \times 100$$

$$C_s^{\text{occasions}} = \frac{O_s^{\text{click}}}{O_s} \times 100$$

KPI 5 - Hallucination rate (AI systems) / rate of unreliable results (traditional search). The former is the proportion of generated statements or references not supported by the source document or factual record; the latter is the proportion of queries in set F (queries with a given formulation, e.g., the title or URL of the misinformation article) that return at least one unreliable result in the top ten ($r \leq 10$), with the reliability of source d measured by its NewsGuard score $NG(d)$ and a threshold of 60 (Aslett et al.) (Equation 5).

Equation 5 - Formalisation of KPI 5 (hallucination rate and rate of unreliable results, conditional on query formulation F)

$$H = \frac{n_{\text{hallucinated}}}{n_{\text{total}}} \times 100$$

$$R(F) = \frac{\text{card}\{q \in F : \exists r \leq 10, NG(d_{qr}) < 60\}}{\text{card}\{q \in F\}} \times 100$$

The binary form of KPI 2 is used for studies with a dichotomous decision (Spatharioti et al., Kaiser et al.), the continuous form for Xu et al.; for KPI 5, proportions computed on manually verified counts are used. This heterogeneity motivates the common effect metric introduced below.

For continuous indicators (KPI 1 and KPI 3), the log response ratio (lnRR; Hedges, Gurevitch & Curtis, 1999) is used, with percentage change equal to $(\exp(\ln RR) - 1) \times 100$. For indicators expressed as proportions (KPI 2 and KPI 4), the odds ratio (OR) is used and converted into Cohen's d following Chinn (2000) (Equation 6); for accuracy expressed as a continuous

score (Xu et al.), the lnRR is used. The lnRR and d are not on the same scale, because the conversion would require standard deviations not reported in the primary studies: magnitudes should be compared only within each metric. For Spatharioti et al. the KPI 1 and KPI 3 values are model estimates, for Xu et al. arithmetic means. The estimates, computed from published aggregate values, are reported in Table 9 (Section 4).

Equation 6 - Common effect metric (log response ratio, odds ratio and conversion to Cohen's d)

$$\begin{aligned}\ln\text{RR} &= \ln\left(\frac{\bar{X}_{\text{AI}}}{\bar{X}_{\text{trad}}}\right) \\ \text{OR} &= \frac{p_{\text{AI}}/(1 - p_{\text{AI}})}{p_{\text{trad}}/(1 - p_{\text{trad}})} \\ d &= \ln(\text{OR}) \cdot \frac{\sqrt{3}}{\pi}\end{aligned}$$

The effect measures in Section 3 and Table 9 are descriptive synthesis estimates, not meta-analytic ones: no pooled weighted effect is computed. The estimates are not independent, because several comparisons come from the same sample (the three tasks of Xu et al. performed by the same 95 participants; the two experiments of Spatharioti et al.; the two domains of Kaiser et al., conditions of a single factorial design). Counts of comparisons should therefore be read as multiple measures rather than independent replications, confidence intervals computed under independence understate uncertainty, and the strength of evidence in Table 10 is expressed qualitatively. In the absence of individual data, this dependence is not modelled with a hierarchical effects model.

Part of the tests in Section 3 is computed on counts reconstructed from rounded percentages; for Kaiser et al. (2025a) the counts are 404/412 and 334/359 (smartphones), 225/382 and 145/373 (theme parks), 95/794 and 534/732 (external sources). The robustness analysis repeats the calculations shifting the counts of the two conditions by one unit in the direction that reduces the size of the effect («unfavourable» scenario) or in the opposite direction («favourable» scenario), and recomputes χ^2 (without continuity correction), OR, the 95% confidence interval of ln(OR) and d following Chinn (2000) (Table 9b, Section 4).

In all scenarios the sign of the effect, its order of magnitude and its statistical significance ($p < 0.01$) remain unchanged. The conventional magnitude class is also unchanged, with one exception: in the smartphone domain the favourable scenario moves d from 0.73 to 0.83, just above the large-effect threshold ($|d| \geq 0.80$), i.e., towards a larger effect. The conclusions of Sections 3.2 and 3.4 therefore do not depend on the

reconstruction adopted. For the difficult task of Spatharioti et al. the perturbation is not applicable, because the published values are probabilities estimated from the authors' model.

3. Results: Key Performance Indicators of search behaviour

3.1. KPI 1 - Number of queries per task

In Xu et al. (2023) the number of queries does not differ significantly over the three tasks combined (7.36 for ChatGPT vs. 8.13 for Google), but the effect depends on task complexity: ChatGPT requires fewer queries in the factual task (1.55 vs. 2.13; $p < 0.01$) and, marginally, in the list-compilation task (1.30 vs. 1.65; $p < 0.10$), whereas it requires slightly more in fact-checking (4.51 vs. 4.35; n.s.). Spatharioti et al. (2025) confirm the overall direction: 1.9 queries per task (95% CI: 1.7–2.2) vs. 2.5 (95% CI: 2.1–3.0), a modest but significant difference ($p = 0.02$; Poisson generalised linear model, $z = -2.24$). In relative terms, the reduction is 24% in Spatharioti et al. and 27% in the factual task of Xu et al., whereas the sign reverses in fact-checking (+4%) (Table 4).

The observational studies measure related but distinct quantities. Shi et al. (2026) estimate that expanding access to ChatGPT Search (intention-to-treat effect) reduces the volume of queries to traditional search engines by 9.4% on average, rising to 17.0% after twenty weeks of exposure, consistent with a gradual change in search habits. Iannelli and Ai (2026; opt-in panel, United States and United Kingdom, February 2026, replicated on March 2026 data) show that the assistant does not uniformly compress search into a single interaction: in 37.1% of sessions web activity alternates with use of the assistant, in 10.5% it only follows it (AI-first) and in 18.3% it only precedes it (AI-last); the share of sessions without any transition to the external web is discussed under KPI 4. Values from version v1 of the preprint are not used.

Table 4: Mean number of queries per task in the two experimental studies

Study	Task/domain	AI system	Traditional search	Difference	Sig.
Xu et al. (2023)	Task 1 — factual question	1.55	2.13	-0.58	$p < 0.01$
Xu et al. (2023)	Task 2 — list of sites/links	1.30	1.65	-0.35	$p < 0.10$
Xu et al. (2023)	Task 3 — fact-checking	4.51	4.35	+0.16	n.s.
Xu et al. (2023)	Total, 3 tasks	7.36	8.13	-0.77	n.s.
Spatharioti et al. (2025)	Product decision (mean)	1.9	2.5	-0.6	$p = 0.02$

Source: Xu, R. et al. (2023). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience; Spatharioti, S. E. et al. (2025). Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In Proceedings of the 2025

CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan. ACM.

3.2. KPI 2 - Task success rate (accuracy)

In Xu et al. (2023), overall accuracy does not differ significantly between ChatGPT and Google Search (85.5% vs. 87.7%; n.s.), but the breakdown by task reveals opposite patterns: ChatGPT achieves a full score in the factual task (100.0% vs. 81.9%; $p<0.01$), comparable results in the list of sites under the general criterion (98.1% vs. 97.4%; n.s.), but lower results under the specific criterion (50.0% vs. 88.8%; $p<0.01$) and in fact-checking (58.3% vs. 83.7%; $p<0.01$).

The authors attribute the decline to ChatGPT uncritically reproducing inaccuracies contained in the user's prompt; 70.8% of participants in the ChatGPT group showed overreliance. In Spatharioti et al. (2025), accuracy is comparable in the four tasks in which the AI system provides correct information (95.3% vs. 92.3%; n.s.), whereas in the fifth task, designed so that the system would provide incorrect information, it falls to 47% vs. 93% (probabilities estimated from the authors' model; $p<0.001$): 30 of the 51 participants in the AI condition stopped after a single query and 23 of them (77%) chose the wrong option (Table 5).

Table 5: Task success rate/accuracy in Xu et al. (2023) and Spatharioti et al. (2025)

Study	Task	AI system	Traditional search	Sig.
Xu et al. (2023)	Task 1 - factual (score %)	100.0%	81.9%	$p<0.01$
Xu et al. (2023)	Task 2 - list of sites, general criterion (score %)	98.1%	97.4%	n.s.
Xu et al. (2023)	Task 2 - list of sites, specific criterion (score %)	50.0%	88.8%	$p<0.01$
Xu et al. (2023)	Task 3 - fact-checking (score %)	58.3%	83.7%	$p<0.01$
Xu et al. (2023)	Mean, 3 tasks (score %)	85.5%	87.7%	n.s.
Spatharioti et al. (2025)	Easy tasks 1-4 (% correct decisions)	95.3%	92.3%	n.s.
Spatharioti et al. (2025)	Difficult task 5 (% correct decisions)	47.0%	93.0%	$p<0.001$

Source: Xu, R. et al. (2023). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience; Spatharioti, S. E. et al. (2025). Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan. ACM.

In the second experiment, colour highlighting based on token generation probability more than doubled accuracy in the difficult task, from 26% (control condition) to 53-58%, increasing decision time without changing user satisfaction; the 26% is not comparable with the 47% of the

first experiment (GPT-3, three tasks instead of five). Highlighting acts as a trust-calibration tool (Lee & See, 2004).

Kaiser et al. (2025a) add a note of caution: in the theme-park task, 59% of ChatGPT users (n=382) identified the correct option, vs. 39% of Google users (n=373); a secondary check on reconstructed counts (225/382 and 145/373) yields $\chi^2(1) = 30.29$, $p < 0.001$, a difference of 20.0 percentage points (approximate 95% CI: 13.0–27.0) and OR = 2.25 (approximate 95% CI: 1.68–3.02). In the smartphone domain, accuracy is high in both conditions (98% with ChatGPT, n=412; 93% with Google, n=359; 404/412 and 334/359: $\chi^2(1) = 11.81$; $p < 0.001$; OR = 3.78, approximate 95% CI: 1.68–8.49).

This apparent contradiction indicates that the sign of the effect on accuracy is not a stable property of the AI system but depends on the correctness of the generated content: when the output is correct, the system's synthesis outperforms scrolling through a list of links (Kaiser et al.); when it is wrong, user trust leads to a collapse in accuracy (Spatharioti et al.). In standardised terms (Table 9), the sign moves from $d = 0.45$ in favour of the AI system (Kaiser et al.) to $d = -1.49$ against it (Spatharioti et al., task with induced error).

3.3. KPI 3 - Total time to task completion

In Xu et al. (2023), system logs show a reduction in time in favour of ChatGPT in all three tasks ($p < 0.01$ for each task): 5.79 vs. 14.95 minutes in total (-61.3%); self-reported time confirms the direction (11.35 vs. 18.75 minutes; -39.5%). In Spatharioti et al. (2025), the estimated mean time is 1.6 minutes (95% CI: 1.4–1.9) vs. 3.4 minutes (95% CI: 2.8–4.1), a reduction of about 53% (linear mixed model on log-transformed task time; $p < 0.001$) (Table 6).

Table 6: Total time to task completion in the two experimental studies (for Xu et al., system-log times)

Study	Task	AI system	Traditional search	Relative Δ	Sig.
Xu et al. (2023)	Task 1	0.34 min	2.42 min	-86.0%	$p < 0.01$
Xu et al. (2023)	Task 2	0.52 min	2.78 min	-81.3%	$p < 0.01$
Xu et al. (2023)	Task 3	4.93 min	9.81 min	-49.7%	$p < 0.01$
Xu et al. (2023)	Total, 3 tasks	5.79 min	14.95 min	-61.3%	$p < 0.01$
Spatharioti et al. (2025)	Mean decision task	1.6 min	3.4 min	-52.9%	$p < 0.001$

Source: Xu, R. et al. (2023). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience; Spatharioti, S. E. et al. (2025). Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan. ACM.

Note: for Xu et al., self-reported times show the same direction (Task 1: 1.83 vs. 3.37 min; Task 2: 2.40 vs. 3.61 min; Task 3: 7.12 vs. 11.78 min; total: 11.35 vs. 18.75 min; all differences $p < 0.01$).

In the fact-checking task of Xu et al., the time saving (-49.7%) coincides with the drop in accuracy (-25.4 percentage points), and in Spatharioti et al. most wrong answers come from participants who stopped after a single query: part of the time reduction may reflect less verification of information rather than pure cognitive efficiency, a hypothesis to be tested on individual-level data.

Time per query (Equation 3) falls from 1.84 to 0.79 minutes in Xu et al. (-57%, system logs) and from 1.36 to 0.84 minutes in Spatharioti et al. (-38%), reductions larger than those in the number of queries (-9% and -24% respectively). On the log scale, $\ln RR(T) = \ln RR(Q) + \ln RR(E)$: in Xu et al. $-0.95 = -0.10 - 0.85$, so the reduction in the number of queries accounts for about 10% of the overall time reduction; in Spatharioti et al. $-0.75 = -0.27 - 0.48$, a share of about 36%. The saving therefore stems mainly from the compression of each interaction, consistent with information foraging theory (Pirolli & Card, 1999) and with cognitive offloading (Risko & Gilbert, 2016).

3.4. KPI 4 - Rate of consultation of external sources (proxy for cross-verification)

Outbound clicks are an observable proxy for, not a direct measure of, cross-verification (RQ4). In Kaiser et al. (2025a), only 12% of the 794 users assigned to ChatGPT visited at least one external website, vs. 73% of the 732 users assigned to Google Search (95/794 and 534/732: $\chi^2(1) = 584.66$; $p < 0.001$; OR = 0.05, approximate 95% CI: 0.04–0.07). In Shi et al. (2026), ChatGPT generates an outbound click in 5.2% of conversation sessions, vs. 31.1% of Google queries; since the unit of analysis differs, the comparison is descriptive and, if anything, understates the gap per individual request (Table 7).

Table 7: Rate of consultation of external sources/outbound clicks in the two studies considered

Study	Method	AI system	Traditional search	Unit of analysis / note
Kaiser et al. (2025a)	Randomised online experiment, N=1 526	12%	73%	% of users; $\chi^2(1) = 584.66$, $p < 0.001$; counts reconstructed from rounded percentages
Shi et al. (2026)	Real clickstream (Comscore panel, desktop only), N=45 386 households	5.2%	31.1%	% of sessions (ChatGPT) vs % of queries (Google); descriptive comparison

Source: Kaiser, C. et al. (2025a). How Generative AI is Transforming Consumer Decision-Making. NIM INSIGHTS Research Magazine, Vol. 7. Nuremberg Institute for Market Decisions (NIM); Shi, Q. et al. (2026). Answering Without Referring: How AI Search Rewrites the Web's Economic Bargain. arXiv:2607.07652.

Using the same unit of analysis for both systems, Iannelli and Ai (2026) find that 34.1% of sessions with an AI assistant (95% CI: 33.5–34.7) involve no transition to the external web, vs. 19.5% of sessions centred on traditional search by the same users (+13.0 percentage points; 95% CI: 12.5–13.6).

The convergence of experimental and observational evidence is consistent with the overreliance suggested by Xu et al. (2023) and Spatharioti et al. (2025) and with automation complacency (Parasuraman & Manzey, 2010).

A system that appears reliable reduces the monitoring of its outputs (Lee & See, 2004), and response fluency, interpreted as a cue of truthfulness (Alter & Oppenheimer, 2009), fosters heuristic rather than systematic processing (Chaiken, 1980). When they occur, ChatGPT's outbound clicks go to different destinations, with a markedly lower share of advertising-supported websites (-27.6 percentage points compared with Google; Shi et al., 2026).

3.5. KPI 5 - Hallucination rate (AI systems) / rate of unreliable results (traditional search)

The hallucination rate (generation of incorrect content) and the rate of unreliable results of a search engine (exposure to pre-existing incorrect content) are distinct phenomena, although they share the same practical outcome for the user; they are therefore presented as complementary evidence rather than as a symmetric comparison (RQ5).

The literature further distinguishes between faithfulness and factuality hallucinations (Ji et al., 2023; Huang et al., 2025): the Vectara Hallucination Leaderboard measures the former, Chelli et al. (2024) and Dahl et al. (2024) the latter.

On the Vectara Hallucination Leaderboard (HHEM-2.3, update of 11 May 2026), an industry benchmark not subject to peer review and based on a grounded summarisation task, the best model has a rate of 1.8% and the other top models range between 3% and 4.1%.

In open-ended factual retrieval the values rise sharply: Chelli et al. (2024) report rates of 39.6% for GPT-3.5, 28.6% for GPT-4 and 91.4% for Bard in generating medical bibliographic references ($p < 0.001$); Dahl et al. (2024) report 58% for ChatGPT-4 and 88% for Llama 2 on US federal court cases.

For traditional search, Aslett et al. (2024) show that searching online to evaluate a false news story increases its perceived veracity in all five

experiments. When the query reproduces the title or URL of the misinformation article, 77% of searches return at least one unreliable result in the top ten (NewsGuard score below 60), vs. 21% of other searches ($p < 0.0001$).

In the illustrative case of queries placing the term “engineered” before “famine”, the share is 63% vs. 0%. The authors interpret the phenomenon as a “data void”: information niches in which the only available sources are low-quality ones (Table 8).

Table 8: Hallucination rate of AI systems and rate of unreliable results in Google Search, by source and domain

Source	System/domain	Error rate	Nature of the data
Vectara HHEM Leaderboard	Top-tier LLMs, grounded summarisation	1.8–4.1%	Standardised benchmark, constrained (grounded) task
Chelli et al. (2024), JMIR	GPT-3.5 / GPT-4 / Bard, medical bibliographic references	39.6% / 28.6% / 91.4%	Peer-reviewed study, $p < 0.001$
Dahl et al. (2024)	ChatGPT-4 / Llama 2, US court cases	58% / 88%	Peer-reviewed study, high-risk domain
Aslett et al. (2024), Nature	Google Search, queries without the false article's title/URL	21%	Peer-reviewed study, 5 experiments, $N > 14\,000$; at least one unreliable result in the top 10
Aslett et al. (2024), Nature	Google Search, queries with the false article's title/URL	77%	Same study, “data void” condition (illustrative example “engineered famine”: 63% vs. 0%; not the main result)

Source: Vectara Hallucination Leaderboard (HHEM); Chelli et al. (2024), Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. JMIR, 26; Dahl et al. (2024), Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis; Aslett et al. (2024), Online searches to evaluate misinformation can increase its perceived veracity. Nature, 625(7995).

The risk of error therefore depends on the context of use in both paradigms: on grounding for AI systems and on query formulation for traditional search. The two quantities are not directly comparable (presence of at least one unreliable result in the top ten, on the one hand; share of wrong answers, on the other), but they indicate that source verification remains a critical user competence, regardless of the tool used.

4. Discussion

Taken together, the five KPIs (Figs. 9 and 10) provide a consistent but not uniform picture. The most robust result concerns time (KPI 3, RQ3): the AI system reduces it in all available comparisons ($p < 0.01$), with reductions between 49.7% and 86.0% per task in Xu et al. and of 53% in Spatharioti et al. The effect on the number of queries (KPI 1, RQ1) is weaker

and task-dependent (-27% in the factual task of Xu et al.; -24% in Spatharioti et al.; +4%, n.s., in the fact-checking task of Xu et al.).

Table 9: Common effect metric: relative change (lnRR) and odds ratio (OR) of the AI system relative to traditional search

KPI	Study and comparison	AI system	Traditional search	Effect (AI vs. traditional)
KPI 1	Xu et al. (2023), task 1	1.55	2.13	lnRR = -0.32 (-27%)
KPI 1	Xu et al. (2023), task 3	4.51	4.35	lnRR = +0.04 (+4%)
KPI 1	Xu et al. (2023), total	7.36	8.13	lnRR = -0.10 (-9%)
KPI 1	Spatharioti et al. (2025)	1.9	2.5	lnRR = -0.27 (-24%; 95% CI: -39%/-5%)
KPI 2	Xu et al. (2023), task 1 factual (score)	100.0%	81.9%	lnRR = +0.20 (+22%)
KPI 2	Xu et al. (2023), task 3 fact-checking (score)	58.3%	83.7%	lnRR = -0.36 (-30%)
KPI 2	Spatharioti et al. (2025), easy tasks	95.3%	92.3%	OR = 1.69; d = 0.29
KPI 2	Spatharioti et al. (2025), difficult task	47%	93%	OR = 0.07 (95% CI: 0.02–0.29); d = -1.49 (95% CI: -2.31/-0.68)
KPI 2	Kaiser et al. (2025a), theme parks	59%	39%	OR = 2.25 (95% CI: 1.68–3.02); d = 0.45 (95% CI: 0.29/0.61)
KPI 2	Kaiser et al. (2025a), smartphone	98%	93%	OR = 3.78 (95% CI: 1.68–8.49); d = 0.73 (95% CI: 0.29/1.18)
KPI 3	Xu et al. (2023), total (log)	5.79 min	14.95 min	lnRR = -0.95 (-61%)
KPI 3	Spatharioti et al. (2025)	1.6 min	3.4 min	lnRR = -0.75 (-53%; 95% CI: -63%/-40%)
KPI 4	Kaiser et al. (2025a)	12%	73%	OR = 0.05 (95% CI: 0.04–0.07); d = -1.65 (95% CI: -1.80/-1.50)

Source: authors' calculations based on data from Xu et al. (2023), Spatharioti et al. (2025) and Kaiser et al. (2025a). lnRR = $\ln(\text{AI mean} / \text{traditional mean})$, with the percentage change in brackets; $d = \ln(\text{OR}) \times \sqrt{3/\pi}$ (Chinn, 2000). The confidence intervals of the lnRR for Spatharioti et al. are approximated from the intervals of the two groups on the log scale, assuming independence. The confidence intervals of the ORs are computed from the standard error of $\ln(\text{OR})$: for Kaiser et al. with the actual cell sizes (382 and 373 in the theme-park domain, 412 and 359 in the smartphone domain, 794 and 732 per tool), with ORs, intervals and χ^2 tests computed on whole counts reconstructed from rounded percentages (e.g., 404/412 and 334/359 in the smartphone domain), and therefore approximate; for the difficult task of Spatharioti et al. the percentages are probabilities estimated from the authors' model (51 participants in the AI condition, out of 90 recruited; the authors do not report raw counts for this task) and the standard error of $\ln(\text{OR})$ is derived from the confidence intervals of these probabilities on the logit scale, assuming

independence between groups; the confidence intervals of d are obtained by applying the Chinn (2000) conversion to the bounds of the $\ln(\text{OR})$ interval; for the easy tasks of Spatharioti et al. no interval is computed, because repeated decisions by the same participant are not independent. For Xu et al. accuracy is a score rather than a proportion, so the $\ln\text{RR}$ is reported instead of the OR; for Shi et al. (2026) no measure is computed because of the different unit of analysis between the two systems.

Table 9b: Robustness analysis of reconstructed counts for Kaiser et al. (2025a)

Comparison	Scenario	OR	95% CI	d	$\chi^2(1)$
Kaiser — smartphone	unfavourable	3.21	1.47–7.00	0.64	9.49
	central	3.78	1.68–8.49	0.73	11.81
	favourable	4.52	1.94–10.54	0.83	14.39
Kaiser — theme parks	unfavourable	2.20	1.65–2.95	0.44	28.71
	central	2.25	1.68–3.02	0.45	30.29
	favourable	2.30	1.72–3.08	0.46	31.91
Kaiser — external sources	unfavourable	0.051	0.039–0.067	-1.64	579.64
	central	0.050	0.038–0.066	-1.65	584.66
	favourable	0.049	0.038–0.065	-1.66	589.71

Note: unfavourable scenario = shift of one unit per condition in the direction that reduces the size of the effect; favourable scenario = opposite shift. χ^2 without continuity correction; 95% CI computed from the standard error of $\ln(\text{OR})$; $d = \ln(\text{OR}) \times \sqrt{3/\pi}$ (Chinn, 2000). Source: authors' calculations based on data from Kaiser et al. (2025a).

In practical terms, the most relevant finding concerns accuracy (KPI 2, RQ2): efficiency translates into correct decisions only when the output is correct. When an error is present, time is still reduced (-49.7% in the fact-checking task of Xu et al.), but accuracy falls by between 25.4 percentage points (Xu et al.: 58.3% vs. 83.7%) and 46 percentage points (Spatharioti et al.: 47% vs. 93%).

The common mechanism is overreliance, observed in 70.8% of participants in the ChatGPT group in Xu et al. and in the 23 out of 30 participants in Spatharioti et al. who stopped after a single query and chose the wrong option. Uncertainty highlighting shows that this trade-off can be mitigated through design choices aimed at calibrating user trust (Lee & See, 2004).

The comparison between Kaiser et al. (2025a) and Spatharioti et al. (2025) indicates that the sign of the effect on accuracy depends, case by case, on the correctness of the generated content. KPI 4 (RQ4) provides indirect behavioural evidence of this, consistent with automation complacency: 12% vs. 73% of users in Kaiser et al. ($d = -1.65$) and 5.2% of ChatGPT sessions vs. 31.1% of Google queries in Shi et al. (2026). In spontaneous behaviour, Iannelli and Ai (2026) observe a bifurcation: 34.1% of sessions with an assistant end without any transition to the external web, whereas in 37.1% the assistant is interwoven with searches and visits to external websites, which calls for caution in generalising laboratory findings.

Table 10: Evidence matrix by KPI (synthesis following the SWiM approach)

KPI	Direction of effect (AI vs. traditional)	Available evidence	Consistency	Strength
KPI 1 - Queries	Reduction in simple tasks; no difference in verification tasks	4 experimental comparisons (2 studies); 1 observational study on a related measure (query volume)	Partial (task-dependent)	Moderate
KPI 2 - Accuracy	Variable sign: depends on output correctness	7 experimental comparisons (3 studies)	Low	Moderate
KPI 3 - Time	Reduction	4 of 4 comparisons with $p < 0.01$ (2 studies)	High	High
KPI 4 - External sources	Strong reduction	1 experimental study, 1 observational (+1 complementary)	High, with different units of analysis	Moderate
KPI 5 - Error	Substantial risk in both paradigms, under specific conditions of use	3 studies + 1 benchmark	Not assessable (non-comparable measures)	Low

Note: strength of evidence is assessed qualitatively, drawing on the GRADE approach (Guyatt et al., 2008): the starting level is «high» for randomised comparisons and «low» for observational ones, and is downgraded for imprecision (small samples or wide intervals), inconsistency between studies, indirectness (proxies or non-comparable measures) and risk of bias (preprints, conflicts of interest, assumptions introduced by the authors in the calculations). Comparisons are counted per task, excluding aggregate totals, which are not independent of their component tasks. The synthesis follows the SWiM approach (Campbell et al., 2020).

Finally, KPI 5 (RQ5) shows that the risk of error is not exclusive to AI systems. The data do not allow us to establish whether the absolute probability of encountering incorrect information differs between the two paradigms; rather, they suggest that what changes is the user's propensity to notice it, since the synthetic answer of an AI system tends, in the studies examined, to be accepted as formulated.

4.1. Interpretive model and research propositions

The results can be brought together in an interpretive model that integrates information foraging theory (Pirolli & Card, 1999), the literature on trust in automation (Lee & See, 2004; Parasuraman & Manzey, 2010) and models of heuristic information processing (Chaiken, 1980; Alter & Oppenheimer, 2009), articulated in four testable propositions.

P1 – The conversational system reduces time to task completion mainly by compressing the time per interaction, and only secondarily by reducing the number of queries (KPI 1 and KPI 3).

P2 – The reduced cost of exploration is accompanied by lower consultation of external sources (KPI 4), consistent with automation complacency.

P3 – The effect on accuracy (KPI 2) is moderated by output correctness: positive or null when the output is correct, strongly negative when it is wrong.

P4 – Interface cues communicating uncertainty attenuate the negative effect described in P3, at the cost of longer decision times.

KPI 5 lies upstream of the model, as it defines the probability that the output is wrong. The theoretical contribution of the review is to shift the focus from comparing tools to the relationship between efficiency, verification and accuracy.

4.2. Limitations of the included studies and of the review

The included studies have limitations that affect the interpretation of the results. (i) The experiments use GPT-3, GPT-3.5 and GPT-3.5-turbo (and, in strand B, also GPT-4, Bard and Llama 2), which predate systems with retrieval-augmented generation and source citation, whereas the observational data concern more recent products. (ii) The experimental samples (95, 90 and 120 participants) are limited, Kaiser et al. (N = 1 526) cover only two consumer domains, and no study includes open-ended exploratory tasks. (iii) The AI tool of Spatharioti et al. did not retain conversational memory. (iv) The observational studies do not support causal inferences comparable to experimental ones; Iannelli and Ai (2026) are affiliated with an industry company and use an opt-in panel. (v) The evidence concerns US and UK users and English-language tools. (vi) Outbound clicks (KPI 4) are a proxy for verification, and in Shi et al. (2026) the unit of analysis differs between systems (session for ChatGPT, query for Google). (vii) Hallucinations and unreliable results (KPI 5) are distinct phenomena, juxtaposed for illustrative purposes. (viii) Xu et al. (2023), Iannelli and Ai (2026) and Shi et al. (2026) are preprints, and Kaiser et al. is a research report with an extended abstract; the verification of 5 October 2026 confirmed the versions used (Iannelli and Ai: v3 of 25 August 2026; Shi et al.: v1 of 8 July 2026, also deposited on SSRN, abstract no. 7035298), but values may change in later revisions.

The review also has limitations of its own: study selection is purposive rather than systematic; effect measures are computed from aggregate values and, for Kaiser et al. (2025a), from counts reconstructed from rounded percentages; data extraction was performed by a single reviewer; part of the evidence comes from preprints and from a research report not subject to full peer review; and the evidence concerns almost exclusively US and UK users and English-language tools.

Conclusions

Drawing on eight independent studies - three randomised behavioural experiments (Xu et al., 2023; Spatharioti et al., 2025; Kaiser et al., 2025), two large-scale observational analyses of real-world behaviour (Iannelli & Ai, 2026; Shi, Zhu & Gu, 2026) and three studies measuring content reliability (Chelli et al., 2024; Dahl et al., 2024, for AI systems; Aslett et al., 2024, for traditional search) - the comparison between traditional search and conversational AI reveals a consistent behavioural transition: users of AI systems need less time and, in simpler tasks, fewer queries to complete a task (KPI 1 and KPI 3).

This efficiency advantage, however, is neither unconditional nor uniform in sign. It is reinforced when the system correctly synthesises complex information, but collapses, sometimes markedly, when users do not critically check an incorrect answer (KPI 2), a risk consistent with the lower consultation of external sources observed with AI systems (KPI 4, proxy for cross-verification) and further framed by the hallucination/reliability indicator (KPI 5).

For research and for the practical design of hybrid search tools, the five KPIs analysed here - number of queries, success rate/accuracy, time to completion, rate of consultation of external sources and hallucination/reliability rate - should not be interpreted in isolation but as a system of interdependent indicators, in which gains in behavioural efficiency are always assessed together with their effect on the final accuracy of the user's decision and on the user's propensity to verify the information received independently.

Future research should replicate this comparison with the latest models equipped with integrated web search (e.g., hybrid tools combining document retrieval and conversational synthesis), extend the analysis to more open-ended and exploratory search tasks, and investigate, with dedicated experimental designs, the individual determinants (education, familiarity with artificial intelligence, risk propensity) of the consultation of external sources and of actual verification, measured directly rather than through clicks alone (KPI 4). An experimental design measuring the hallucination rate (KPI 5) and the rate of unreliable results on the same set of queries for both tools would make the comparison statistically symmetric.

Brand strategies are changing accordingly. Brands now seek visibility within AI, since the goal is no longer only the first blue link but being cited as an authoritative source within AI-generated summaries; they also favour topic-cluster content structures and a strong brand reputation beyond search engines (through social media and digital PR). Moreover, LLMs are continuously updated and use the context of queries, previous interactions,

preferences, location and browsing history to deliver highly personalised, context-sensitive search experiences.

In today's digital landscape, therefore, indexing alone is no longer sufficient to remain competitive. Firms need advanced omnichannel strategies based on consistent, synchronised data across all platforms and on equally synchronised, transmedia customer service - prompt, responsive and able to anticipate complaints across store, mobile, desktop and social media. Finally, firms should exploit the data already at their disposal, since initiatives based on its advanced analysis allow them to understand preferences and behaviours and to intercept needs and desires more precisely.

Conflict of Interest: The authors reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The authors did not obtain any funding for this research.

References:

1. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5–16). ACM. <https://doi.org/10.1145/3637528.3671900>
2. Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13(3), 219–235. <https://doi.org/10.1177/1088868309341564>
3. Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J., & Tucker, J. A. (2024). Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625(7995), 548–556. <https://doi.org/10.1038/s41586-023-06883-y>
4. Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <https://doi.org/10.1002/9780470743386>
5. Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, 368, l6890. <https://doi.org/10.1136/bmj.l6890>
6. Caramancion, K. M. (2024). *Large language models vs. search engines: Evaluating user preferences across varied information*

- retrieval scenarios (arXiv:2401.05761). arXiv.
<https://arxiv.org/abs/2401.05761>
7. Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5), 752–766. <https://doi.org/10.1037/0022-3514.39.5.752>
 8. Chapekis, A., & Lieb, A. (2025, July 22). *Google users are less likely to click on links when an AI summary appears in the results*. Pew Research Center. <https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/>
 9. Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9(2), 129–136. [https://doi.org/10.1016/0040-5809\(76\)90040-X](https://doi.org/10.1016/0040-5809(76)90040-X)
 10. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J. L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26, e53164. <https://doi.org/10.2196/53164>
 11. Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19(22), 3127–3131. [https://doi.org/10.1002/1097-0258\(20001130\)19:22<3127::AID-SIM784>3.0.CO;2-M](https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M)
 12. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64–93. <https://doi.org/10.1093/jla/laae003>
 13. Fishkin, R. (2024, July 1). *2024 zero-click search study: For every 1,000 US Google searches, only 374 clicks go to the open web. In the EU, it's 360*. SparkToro. <https://sparktoro.com/blog/2024-zero-click-search-study-for-every-1000-us-google-searches-only-374-clicks-go-to-the-open-web-in-the-eu-its-360/>
 14. Forrester. (2024). *Buyers' journey survey, 2024*. Forrester Research.
 15. G2. (2025). *2025 buyer behavior report*. G2 Research. <https://research.g2.com/cm05-2025-buyer-behavior-report-research-g2>
 16. Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650), 924–926. <https://doi.org/10.1136/bmj.39489.470347.AD>

17. Harsel, L. (2025, July 30). *Google AI Mode's early adoption and SEO impact*. Semrush. <https://www.semrush.com/blog/google-ai-mode-seo-impact/>
18. Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, 80(4), 1150–1156. [https://doi.org/10.1890/0012-9658\(1999\)080\[1150:TMAORR\]2.0.CO;2](https://doi.org/10.1890/0012-9658(1999)080[1150:TMAORR]2.0.CO;2)
19. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. <https://doi.org/10.1145/3703155>
20. Iannelli, M., & Ai, A. (2026). *The new shape of search: How conversational AI recomposes information seeking* (arXiv:2607.04282v3, version of 25 August 2026; v1 of 5 July 2026). arXiv. <https://arxiv.org/abs/2607.04282>
21. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
22. Kaiser, C., Kaiser, J., Schallner, R., & Schneider, S. (2025a). How generative AI is transforming consumer decision-making. *NIM Insights Research Magazine*, 7. Nuremberg Institute for Market Decisions.
23. Kaiser, C., Kaiser, J., Schallner, R., & Schneider, S. (2025b). A new era of online search? A large-scale study of user behavior and personal preferences during practical search tasks with generative AI versus traditional search engines. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. ACM. <https://doi.org/10.1145/3706599.3720123>
24. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
25. Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, 80(6), 69–96. <https://doi.org/10.1509/jm.15.0420>
26. Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7001–7025). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.467>

27. Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), 41–46. <https://doi.org/10.1145/1121949.1121979>
28. Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum*, 55(1), 1–27. <https://doi.org/10.1145/3476415.3476428>
29. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
30. Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, 106(4), 643–675. <https://doi.org/10.1037/0033-295X.106.4.643>
31. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
32. Shah, C., & Bender, E. M. (2022). Situating search. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)* (pp. 221–232). ACM. <https://doi.org/10.1145/3498366.3505816>
33. Shi, Q., Zhu, K., & Gu, K. (2026). *Answering without referring: How AI search rewrites the web's economic bargain* (arXiv:2607.07652v1, version of 8 July 2026). arXiv. <https://arxiv.org/abs/2607.07652> (also deposited on SSRN, abstract no. 7035298, <https://doi.org/10.2139/ssrn.7035298>)
34. Spatharioti, S. E., Rothschild, D., Goldstein, D. G., & Hofman, J. M. (2025). Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM. <https://doi.org/10.1145/3706598.3714082>
35. Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213–225. <https://doi.org/10.1086/258464>
36. TrustRadius. (2025). *Bridging the trust gap: B2B tech buying in the age of AI*. <https://go.trustradius.com/rs/827-FOI-687/images/TrustRadius-Bridging-the-Trust-Gap-B2B-Tech-Buying-in-the-Age-of-AI.pdf>
37. Vectara. (2026). *Hallucination leaderboard* (HHEM-2.3, update of 11 May 2026) [GitHub repository]. Retrieved October 3, 2026, from <https://github.com/vectara/hallucination-leaderboard>
38. Xu, R., Feng, Y., & Chen, H. (2023). *ChatGPT vs. Google: A comparative study of search performance and user experience* (arXiv:2307.01135). arXiv. <https://arxiv.org/abs/2307.01135>