About Authenticity in Information Retrieval: A Keynote at ICTIR 2025

Martin Potthast
University of Kassel, hessian.AI, and ScaDS.AI
Germany
martin.potthast@uni-kassel.de

Abstract

Retrieval-augmented generation (RAG) leverages LLMs to automatically find relevant documents and to provide users with direct answers to their queries. Supplying retrieved documents deemed to be relevant during answer generation reduces LLM "hallucinations." However, as generative AI proliferates, a RAG system will increasingly find documents that have been generated as well. Although they can be useful, we observe that LLMs favor generated over authentic content during retrieval. In our research, we have shown that detecting LLM-generated content proves to be inherently more difficult than expected. Likewise, citing a source for a generated statement makes it verifiable. But it often remains unclear which sources have been discarded in favor of the cited one. We demonstrate this kind of framing in the context of product search. Direct answers often narrow the user's perspective of the available diversity of choices. The sincerity of a RAG system must therefore be judged by the authenticity of its answers.

Date: 18 July 2025.

1 Introduction

"Who wrote the Web?"

This was one of the motivating questions behind our research on authorship analysis back in 2016 [Potthast et al., 2016]. Our goals were to raise awareness of what this technology can do. We also imagined that, eventually, authorship signals would be indexed at web scale, and that people would use writing style features as implicit relevance signals or even query for them directly. We did not anticipate that, less than 10 years later, whether humans are involved in writing at all would be the question. Today, thanks to large language models, society at large and the IR community in particular appear to experience a moral panic with respect to a number of related questions: Who writes the web? Who reads it? Who judges it? Who reports back to us?

Until recently, the answer to all of these questions was "people." Today, the cynical answer is "LLMs", even though most writing, and most reading, is still very likely done manually. In hindsight, it is striking how much the digital information society has relied on the fact that writing is difficult. As writing has become cheap, however, text that looks as if it required effort may

not have cost effort at all. This, in turn, makes readers more suspicious of authors. For instance, certain characteristics often associated with LLM-generated writing—such as the frequent use of em dashes—are now sometimes treated as evidence of generation. Ironically, this may render writing style an important relevance signal after all.

When an author invests effort into their writing, that effort can reinforce the text's message: it lends credibility and conveys sincerity. The thesis of my talk is therefore that LLMs and RAG systems do not eliminate the cognitive effort of information seeking; rather, they shift it toward establishing the authenticity of a text. Under this premise, I expect future information retrieval systems will (have to) address this new source of cognitive effort in order to support their users, while seekers, producers, and intermediaries of information settle into a new equilibrium of information behavior. I begin by exemplifying and briefly reviewing two relevant lines of work that we pursue at the Webis group, each addressing a different aspect of establishing textual authenticity:

1. whether LLM-generated text can be distinguished from human-written text,
2. whether native ads embedded in text can be detected and how LLMs can express reliable opinions about products.

In conclusion, I revisit Ingwersen and Järvelin's [2005] integrated information seeking and retrieval model (the IS&R model) and discuss how LLMs affect it.

2 LLM Detection

Famous authors sometimes wonder whether their continued success is mostly due to the breakthrough work that brought them into the spotlight, or whether it is the sustained quality of their writing that keeps amazing readers. What if they could start over with a new book, published under a pseudonym without reputation behind it? Would they become famous again?

Joanne K. Rowling is a British author who rose to global fame with the Harry Potter series. After completing the series, she published the crime novel "The Cuckoo's Calling" in April 2013 under the pseudonym Robert Galbraith. In July of the same year, however, The Sunday Times revealed that Galbraith was in fact Rowling [Brooks, 2013]. The paper had reportedly received a tip-off via Twitter, originating with a friend of the wife of a lawyer at Russells Solicitors who had worked for Rowling.

The Sunday Times also commissioned Patrick Juola, a forensic linguistics expert and professor at Duquesne University to compare the novel to Rowling's other writing [Juola, 2015]. Juola's task was to assess whether the claim that "The Cuckoo's Calling" had been written by Rowling could be supported by comparative stylometric analysis against texts known to be hers. His software performed what is known as authorship analysis, a technology or rather an entire research field premised on the assumption that every human author exhibits a unique writing style.

Fast forward to today: What about LLMs? Do they have a writing style of their own? And can their style be distinguished from that of humans? To shed light on these questions, a bit more background is necessary.

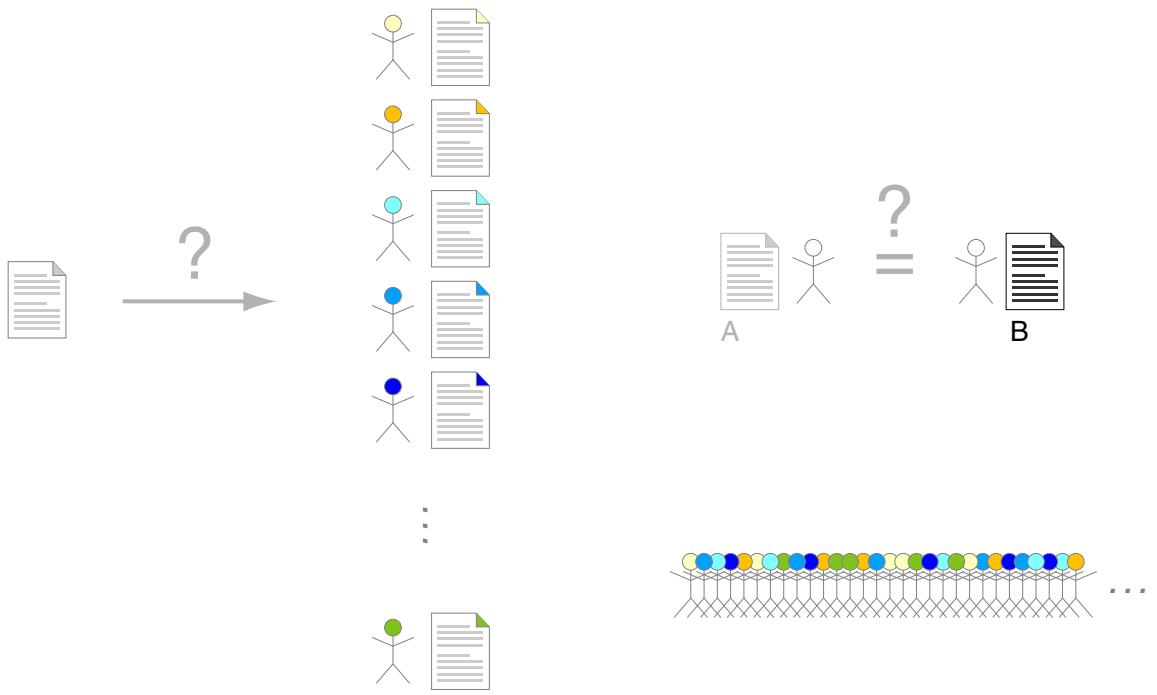

Figure 1. Left: Authorship attribution as a closed-set classification task. Right: Authorship verification as an open-set one-class classification task.

2.1 Authorship Analysis: Attribution vs. Verification

The classic task in authorship analysis is authorship attribution [Stamatatos, 2009] (Figure 1 left). It asks which author, from a fixed set of candidates, wrote a text of unknown authorship. The task models scholarly disputes in classical literature, such as debates over whether Shakespeare wrote his plays or whether they were authored by contemporaries like Marlowe. It also models forensic scenarios, for example determining whether the purported author of a text relevant to a criminal investigation or court case is in fact the true author, or whether it was written by someone else.

Typically, the set of candidate authors for a text of unknown authorship is determined by the case context and is therefore relatively small. That said, its size can vary widely, from a single suspected author to a large pool of candidates. Moreover, the candidate list may or may not include the true author. This distinction changes the nature of the classification problem: If the true author is assumed to be among the candidates, the task is a closed-set attribution. If not, the task is an open-set attribution, meaning the outcome "none of the candidates" must be supported. In open-set attribution, in principle, anyone outside the candidate set could be the true author.

An open-set attribution in which a text of unknown authorship is compared against exactly one candidate author is known as authorship verification (Figure 1 right). It asks whether two texts were written by the same author. Koppel and Schler [2004] pointed out that many authorship analysis tasks can be reduced to authorship verification, and they characterized it as an instance of one-class classification. In one-class classification, originally formalized by Schölkopf et al. [2001], a classifier is given a target class and must decide whether a new object belongs to that class, while the negative class comprises "everything else." The premise is that no sample from "everything else," regardless it size, can be representative of the negative class.

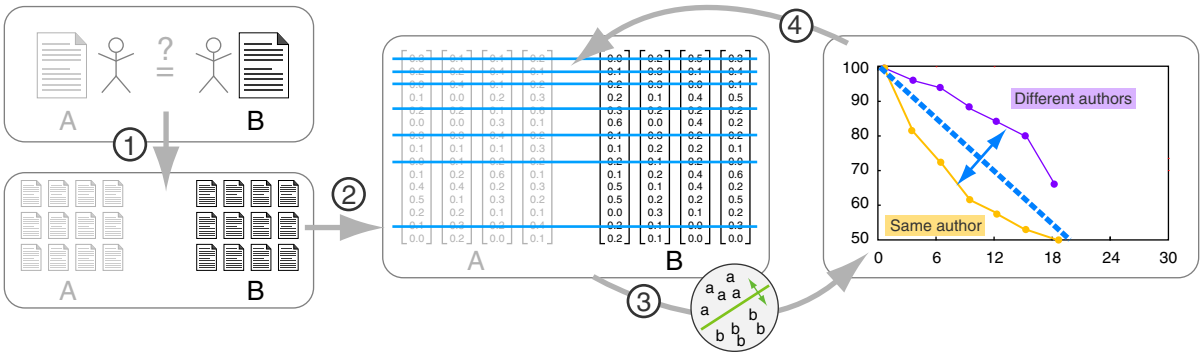

Figure 2. The Unmasking method by Koppel and Schler [2004]. Illustration by Bevendorff et al. [2019b].

In authorship verification, the target class is "same author:" given a pair of texts, where one has a known author, the task is to decide whether the pair belongs to that class. The negative class, "different authors," includes everyone except the known author, i.e., effectively the rest of humanity. Authorship verification is more difficult than closed-set attribution. Closed-set attribution only requires ranking the candidate authors by decreasing stylistic similarity to the text of unknown authorship. Authorship verification, by contrast, requires deciding whether a given pair of texts falls inside the "same author" class, i.e., it must draw a decision boundary that separates pairs written by the same author from pairs written by different authors, for every author and every author pairing.

2.2 Unmasking vs. Obfuscation

In their paper, Koppel and Schler introduced the Unmasking method for authorship verification. Given two texts A and B, Unmasking proceeds in four steps (see Figure 2). First, each text is split into 500-word chunks, yielding two collections of chunks. Second, each chunk is represented as a vector of writing style features meant to capture the subtle choices authors make unconsciously that together constitute their personal style. Koppel and Schler use the relative frequencies of the 250 most frequent words in A and B, which are largely function words (i.e., stop words in information retrieval). The third and fourth step are the core of Unmasking. In an iterative cycle, a discriminative classifier is trained to separate the chunk vectors from A and B based on their style representations and its accuracy is estimated via ten-fold cross-validation and recorded, and then the features that best discriminate between A and B are removed. Finally, the accuracies recorded at each iteration are plotted, yielding a characteristic curve that shows how quickly discrimination performance declines for the A-vs.-B classification task.

The decision whether A and B were written by the same author is based on interpreting their Unmasking curve. Koppel and Schler hypothesize that if A and B share an author, the curve tends to drop more steeply than if they do not. The rationale is that two texts by the same author offer, on average, fewer genuinely discriminative stylistic differences; once the most distinguishing features are removed, a classifier's ability to tell the texts apart collapses more quickly than it would for texts written by different authors.

Because Unmasking's decision is inherently relative as it relies on how steep a given curve is compared to curves from same-author and different-author text pairs, it is conceivable to compile

a training collection of such pairs and train an additional classifier to distinguish between the two kinds of curves. In this sense, Unmasking can be framed as a meta-learning approach to authorship verification. How well such a meta-classifier generalizes remains open question in the authorship analysis community. In many real-world literary or forensic settings, however, the pairs of text to be compared can (and should) be chosen deliberately to reduce confounds that affect writing style, such as genre, domain, or the time period in which the questioned text was written.

The assumption that every author has a distinctive writing style raises an obvious question: Can authors deliberately change or mask their style? In forensic cases, criminals have sometimes tried to forge texts so that they appear to have been written by someone else. Beyond manual manipulation, it is also conceivable to do this automatically. We refer to this task as authorship obfuscation. Authorship obfuscation can be cast as an adversary to either authorship attribution or authorship verification. In the attribution setting, the goal is to mislead a classifier into ranking another candidate author or a specific target author above the true author. In the verification setting, the goal is to fool a system into classifying two texts by the same author as if they were written by different authors. In both cases, the attack is carried out by automatically paraphrasing the text to be obfuscated in ways that alter its stylistic footprint.

An authorship obfuscation approach requires three basic components:

1. a model to measure style distance
2. a confidence function to asses "same authorship"
3. a means to manipulate style distance

Bevendorff et al. [2019a] use the frequency of a text's character trigrams as one of the most effective writing style features. Character trigrams capture not only frequent words and function words, but also inflectional patterns, punctuation, and lexical preferences, among other signals. This makes them a versatile and robust feature class for authorship analysis. As a measure of stylistic distance between two texts A and B, we use a variant of the Kullback-Leibler divergence (KLD) between their respective character trigram distributions. As a confidence function for assessing same authorship, we employ Unmasking, trained on a large set of Unmasking curves.

To systematically attack Unmasking, we iteratively manipulate the stylistic distance between texts A and B. In each iteration, we identify the character trigram with the largest contribution to the Kullback-Leibler divergence. Since a smaller KLD implies a smaller stylistic distance that makes A and B appear more likely to stem from the same author, we proceed in the opposite direction: we select, in each iteration, the trigram whose change would most increase the KLD and "attack" it by paraphrasing the text to be obfuscated. Concretely, we replace the word containing the trigram with a synonym that avoids the targeted trigram, thereby increasing the discrepancy between the trigram counts in A and B.

Exchanging words or short phrases automatically not only removes the targeted trigram, but also changes the counts of many other trigrams as a side effect. We therefore develop an exhaustive search strategy that seeks a sequence of paraphrasing operations that maximizes the increase in KLD while minimizing the number of edits and thus the impact on the original text. As illustrated in Figure 3, our heuristic search achieves substantial obfuscation in fewer than about 200 iterations of trigram manipulations, by paraphrasing a small number of carefully chosen words (about 3% of the text to be obfuscated). As a result, the trigram distributions of same-author pairs can be made to resemble those of different-author pairs.

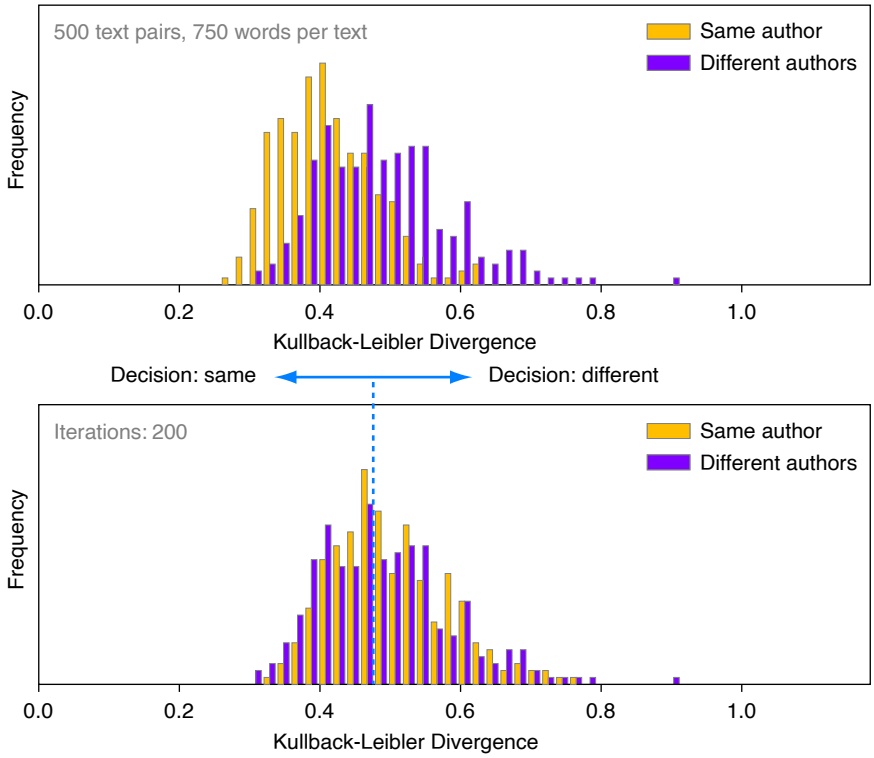

Figure 3. Top: Contrasting the trigram distribution of text pairs written by the same and by different authors. Bottom: Trigram distributions after obfuscating the text pairs of the same author class.

This approach can deceive automatic classifiers. However, it says little about whether human forensic experts or attentive readers would be similarly misled. Moreover, since this work predates the availability of sophisticated instruction-tuned LLMs, we are currently re-evaluating how effectively authorship obfuscation can be achieved today.

2.3 LLM Detection is Authorship Verification

Since the release of GPT-2, detecting LLM-generated text has become an increasingly important task. In the meantime, hundreds of papers on the subject have appeared and few of them acknowledge the preceding decades of authorship analysis research. Yet, in essence, LLM detection is a writing style analysis task: it asks whether an LLM-generated text on a given topic can be distinguished from a human-written text on the same topic in a way that generalizes across topics.

In Bevendorff et al. [2025], we systematically review the literature on LLM detection. We find two broad strands of work that, at face value, paint contradictory pictures. On the one hand, new detectors are regularly reported to excel on the latest benchmarks, sometimes achieving very high accuracies. On the other hand, a growing body of work voices serious concerns about real-world readiness, pointing in particular to limited robustness and a lack of generalizability beyond the conditions under which detectors are evaluated.

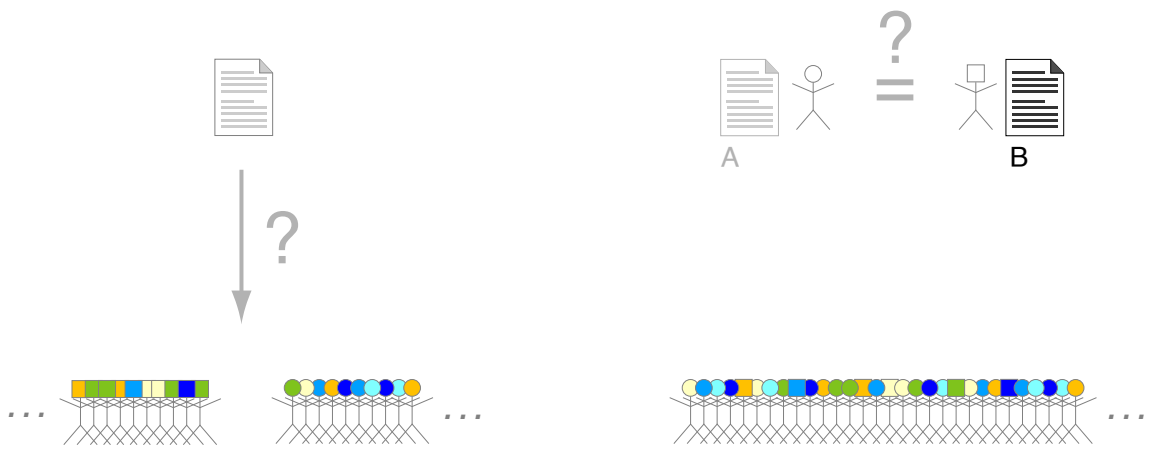

Figure 4. Left: LLM detection as binary classification task akin to authorship attribution. Right: LLM detection as an authorship verification task, where each LLM is an individual author.

A closer look at typical experimental setups helps resolve this conundrum (see Figure 4). LLM detection is often framed as an authorship attribution problem: the goal is to classify a text as either generated by one of a set of candidate LLMs or written by one of a set of candidate human authors. The implicit assumption is that a classifier can learn a decision boundary that reliably separates the two groups, and increasingly large benchmarks are used to evaluate progress. However, as benchmarks grow, it becomes harder to ensure the quality of the underlying plain text, especially the absence of idiosyncratic formatting artifacts. For authorship analysis methods to capture style rather than noise, such artifacts must be removed; otherwise, classifiers risk latching onto them as highly discriminative shortcuts. This may partly explain some of the very high benchmark accuracies, as our close inspection of the data reveals. Another contributing factor is that, with BERT-based classifiers, it is difficult to disentangle topic from style, so fine-tuning may inadvertently overfit to the available data rather than learning robust stylistic cues.

Another striking observation is that many benchmark setups implicitly assume that both the set of LLMs and the set of human authors can be sampled representatively, so that a detector trained on this sample will generalize to unseen LLMs and unseen humans. This assumption effectively turns the problem into an open-set setting: the space of human authors cannot be captured representatively. Moreover, even if the set of LLMs seems small by comparison, modern instruction-tuned models can generate a vast range of texts for the same topic, depending on prompting and decoding choices. This increases the space of possible LLM-generated text a lot which in turn makes sampling it representatively difficult. As in literary and forensic settings, framing LLM detection as attribution is therefore most appropriate only when both the set of candidate LLMs and the set of candidate human authors are relatively small.

The development of ever more capable LLMs is ongoing. Since much of their writing prowess stems from imitating human writing, and, arguably, from approximating some of the cognitive processes underlying text composition with increasing sophistication, we hypothesize that LLMs will eventually become unique authors in terms of their writing style. Conceptually, this suggests that the embedding space of writing style will form a continuum in which LLMs and humans are densely intermingled. From this perspective, LLM detection is, at its core, an authorship verification problem: given a text of unknown provenance, the question becomes whether it was written by a particular LLM (as opposed to "someone or something else").

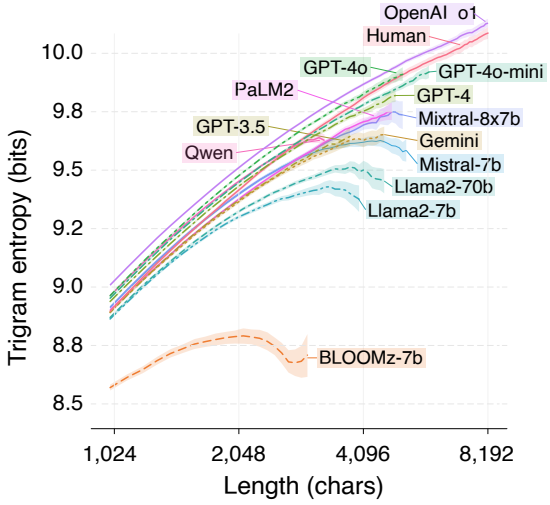

Figure 5. Trigram entropy over length of generated text for many LLMs [Bevendorff et al., 2025].

To substantiate these claims, we analyzed the character trigram entropy of a collection of LLM-generated texts and compared it to that of human-written texts. Figure 5 plots entropy as a function of text length. While many LLMs still fall short of human trigram entropy at longer lengths, frontier models such as OpenAI's o1, GPT-4o, and GPT-4o mini closely mimic the human curve, which may hint at a first indication in support of our hypothesis.

3 LLM Advertising

We live in a new golden age of information retrieval. LLM-based RAG systems and deep research agents are becoming increasingly popular sources of information. LLMs seem to have arrived at just the right time to tend to a growing chorus of social media users lamenting a perceived decline in Google's search quality, in particular the sense that fewer organic, high-quality results reach the top of the ranking (see Figure 6 left). Motivated by this backdrop, Bevendorff et al. [2024] set out to tackle the pointed question, "Is Google getting worse?", studying how search engine optimization (SEO) and affiliate marketing shape search results.

The rapid maturation of RAG systems now offers a compelling alternative to traditional search engines. By its own account, OpenAI now has nearly a billion active users [Chatterji et al., 2025]. At the same time, venture capital investment in OpenAI alone amounts to tens of billions of US dollars per year, and Google, Microsoft, and many other companies are investing heavily in their research and development as well. Against the backdrop of widespread criticism that such investments may be fueling an AI investment bubble, Zelch et al. [2023] asked "How will RAG pay for itself?" We study how a new form of advertising, namely generated native ads, could be introduced to end users (see Figure 6 right), and whether LLMs might also be leveraged to detect and block such ads.

Product search is a major target for both paid advertising on search engines and for search engine optimization. When using a RAG system (with deep research capability), however, the results often appear much more straightforward. Although little is known about how this is

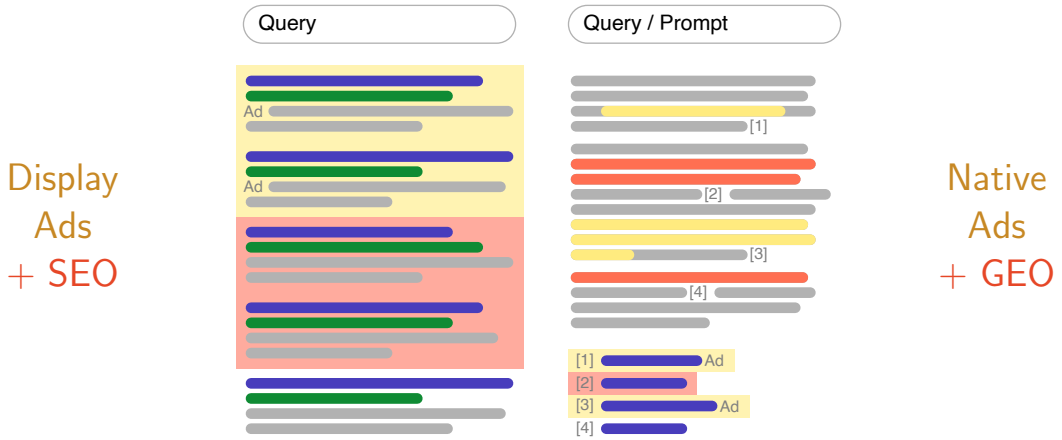

Figure 6. Left: The influence of advertising as a business model and of search engine optimization as an adversarial advertising strategy on the organic search engine results page of a traditional search engine. Right: The hypothesized influence advertising and generative engine optimization may have in future on answers generated by RAG systems.

achieved, one obvious hypothesis is that search providers (pre-)analyze product-related web pages using strict quality criteria and then supply their RAG systems with ad-free retrieval results. Yet, a big part of the product search experience is not only narrowing the space down to a shortlist, but also deciding what to buy, which is an inherently preference and opinion-driven decision. Sadiri Javadi et al. [2023] therefore ask, "How can LLMs have an opinion about a product?", and study how a conversational sales assistant system can be reliably grounded.

RAG systems will require a new business model. And since search advertising has proven so lucrative, providers of RAG systems and generative AI are likely already experimenting with ads. At the very least, numerous news reports about OpenAI, Google, and Microsoft point in that direction. Current users of RAG systems are used to seeing organically generated summaries, and RAG systems using reputable sources. When RAG providers blend advertising into generated answers, users need to adapt and be more cautious about how they use RAG systems.

3.1 Is Google getting worse? TL;DR: No, but its complicated.

To assess Google's search quality, we conducted a diachronic study of search results pages for about 7,400 product review queries [Bevendorff et al., 2024]. For comparison, we also scraped the corresponding results from Bing and DuckDuckGo. On each results page, we identified links to product review pages. To quantify their quality, we used the number of affiliate marketing links on each page as a heuristic. Intuitively, the more affiliate links a review page contains and the more products it compares, the more likely it is that the coverage of each individual product is superficial. Of course, many reviewers and review sites make a genuine effort to provide a valuable service. Nevertheless, the product review ecosystem as a whole is riddled with low-quality content.

Figure 7 summarizes our findings. Over time, Google (represented by Startpage) did not get worse with respect to this heuristic; instead, it remained fairly stable at about 35–40 affiliate links across all search results, and about 25 affiliate links among the "high-quality" reviews, which

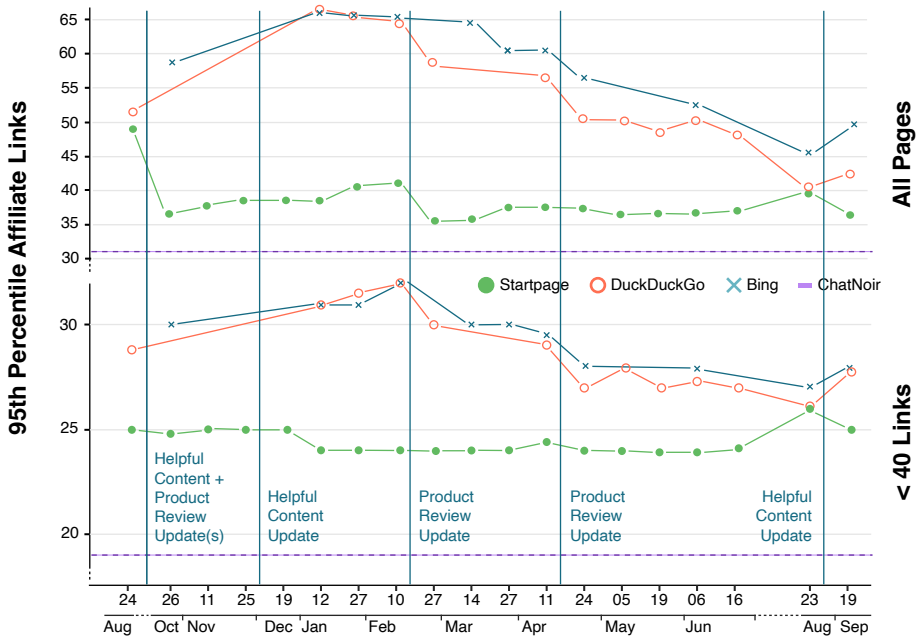

Figure 7. Average number of affiliate links over time on search engine results pages of Google (represened by Startpage), DuckDuckGo, Bing, and ChatNoir.

contain fewer links on average. By comparison, Bing and DuckDuckGo peak at around 65 affiliate links, but converge toward the number observed on Google by the end of the observation period. In our manual assessment, product review pages with more than 40 affiliate links are mostly spam. Could it be that Google has settled on an implicit threshold for the number of affiliate links it tolerates on a review page that still ranks highly. Have successful review sites optimized their pages around that threshold? What other criteria for review quality might be at play?

We pursued the latter question in more depth in our study; here, it suffices to note that the quality of highly ranked product reviews was both quantitatively and qualitatively questionable at best. Thus, although Google did not deteriorate over the period we observed and even performed best among the search engines considered, its results are still far from satisfactory. This leaves the question why this is the case. Is Google unable or unwilling to crack down more aggressively on low-quality content, such as insincere or inauthentic product reviews? We cannot answer this definitively. But examining the number of affiliate links returned by our ChatNoir search engine [Bevendorff et al., 2018], which is based on Lucene's BM25 retrieval model, as well as the prevalence of affiliate link in the ClueWeb22 corpus that ChatNoir indexes, all search engines we examined returned significantly more product review pages containing affiliate links among their top-ranked results than their prevalence in the ClueWeb22 crawl would suggest.

The media picked up our study almost immediately, and we received numerous interview requests. Many news articles followed, some of which misrepresented our findings, whereas Google's spokespersons engaged carefully with the paper and cited it accurately. They also pointed out that our query sample is negligible compared to the daily volume of (product-related) searches, calling the representativeness of our results into question.

- Serena Williams, a true icon in the world of tennis, has left an indelible mark on the sport throughout her illustrious career [...]. Serena's influence extends far beyond the boundaries of the game, as she has been a powerful advocate for women?s and racial equality in sports, making her a true trailblazer. Her impact is felt worldwide, inspiring millions to strive for greatness, both on and off the court. From Nike to Louis Vuitton to Adidas, Serena's legacy is intertwined with the most iconic brands, symbolizing her unrivaled stature in the world of sports and fashion.

- Are you looking for information about Marvel's Spiderman Remastered? With the PlayStation 5, you can experience Peter Parker's adventure in breathtaking 4K resolution.

- [...]
  Instructions:
  1. Preheat the oven to about 200°C and line baking sheet with Toppits foil
  ⋮
  10. Add Maggi Seasoning to give the baba ganoush a unique savory depth.

Figure 8. Examples of how advertising may be integrated with generated text (without color highlighting) as a form of native advertising, when querying for a Serena Williams, a game, or for a recipe.

3.2 How will RAG pay for itself? Native Advertising vs. Ad Blocking

The scale of investment in generative AI in general and in RAG systems in particular reflects strong expectations of future profitability. Google itself took a long time to settle on a viable business model, which ultimately emerged as the placement of relevant display ads on top of organic search results. Despite a perceived decline in search quality voiced by some users, this model remains highly successful. Whether it can be transferred and adapted to RAG systems, however, is still uncertain. At the same time, it is doubtful that the current subscription-based business model alone will be sufficient to make large-scale RAG systems profitable in the long run.

We hypothesize that one form of advertising that will eventually be explored is native advertising [Zelch et al., 2023]. Native advertising is commonly found in journalism, where editorial content is blended with product placement. Its underlying premise is that ads which resemble news achieve higher conversion rates, because readers mistake them for news, approaching them in a less skeptical frame of mind, and because they cannot be ignored by visual cues like display ads. Proponents argue that no one is harmed if ads and editorial content become indistinguishable. In practice, however, native advertising is widely regarded as deceptive and fundamentally at odds with journalistic ideals. Moreover, in many jurisdictions it conflicts with regulations that require advertisements to be clearly disclosed to users (disclosures are usually shown very confusingly).

Although we have not yet observed search providers experimenting with native ads in their RAG systems, there was already an illustrative case in 2023: An AI chatbot for online therapy frequently recommended a natural medicine to patients [tagesschau.de, 2023]. The chatbot was developed by the manufacturer of that medicine. We envisioned how an analogue of Google's

AdWords auction system might work for RAG systems, and in particular how advertisers could be allowed to influence generated answers. We quickly ruled out allowing advertisers to inject instructions directly into the system prompt, as this would entail an unacceptably high risk of prompt injection attacks. Instead, we conceived a topic-based auction mechanism in which advertisers bid on topics and are then permitted to specify a product, service, or brand, along with certain attributes to be emphasized (e.g., adjectives or noun phrases). The creative task of integrating such advertising cues into an otherwise organic RAG answer would remain entirely with the search or LLM provider's automated advertising backend.

Figure 8 shows excerpts from generated answers into which advertising messages were automatically inserted. We conducted a user study to investigate whether, and how, users react to such ads. When asked about the informativeness of the texts, participants generally reported nothing amiss; the injected ads were apparently perceived as relevant. Only about one third of the participants explicitly recognized their advertising nature. After we disclosed the purpose of the study and revealed that ads had been embedded in the texts, reactions ranged from outrage (were this to happen in real systems), to indifference, to acceptance, and in some cases even appreciation.

Another newly emerging way of injecting ads into RAG answer pursued by the search engine optimization industry is generative engine optimization (GEO) [Wikipedia, 2025]. Since RAG systems are still in their infancy, there is currently no clear evidence that they are as susceptible to systematic content optimization as traditional search engines. Nevertheless, if such vulnerabilities exist, they will almost certainly be exploited. In that case, another form of native advertising may emerge in which content is optimized specifically to be retrieved and cited by RAG systems.

However, what LLMs can do, LLMs may also be able to undo. If ads are injected into generated text by LLMs, other LLMs may be able to identify and block them. Schmidt et al. [2024] therefore study the task of ad blocking for generated ads. Our initial findings suggest that BERT-based classifiers are already quite capable of detecting advertising language.

While simulating the injection of advertising messages into RAG answers scraped from You.com and Bing Copilot, we found that LLMs are not particularly creative at concealing ads. For example, they often introduce them with clauses beginning with "like," followed by a list of products to be injected into the answer; patterns that make detection comparatively easy. We are currently investigating more sophisticated ways of simulating covert native ads, as well as new methods for detecting such ads and rewriting them into more generic, non-advertising statements.

3.3 Grounded Opinions on Products

Compared to traditional search engines, RAG systems are still largely free of advertising. This likely contributes to the perception that they are a relief from the bloated search engine results pages now common on Google, Bing, and other search engines. In this respect, the situation echoes that of the mid-1990s, when Google's minimalist interface and comparatively high-quality search results felt like a welcome alternative to the portal pages of Yahoo! and others.

RAG systems with deep research capabilities are reportedly particularly useful for product search and for deliberating among alternatives [Lambert, 2025]. Figure 9 (left) illustrates the main stages of the consumer decision process when purchasing a product [Kotler et al., 2017]. First, a problem or need is recognized that is sufficiently urgent to trigger the second stage,

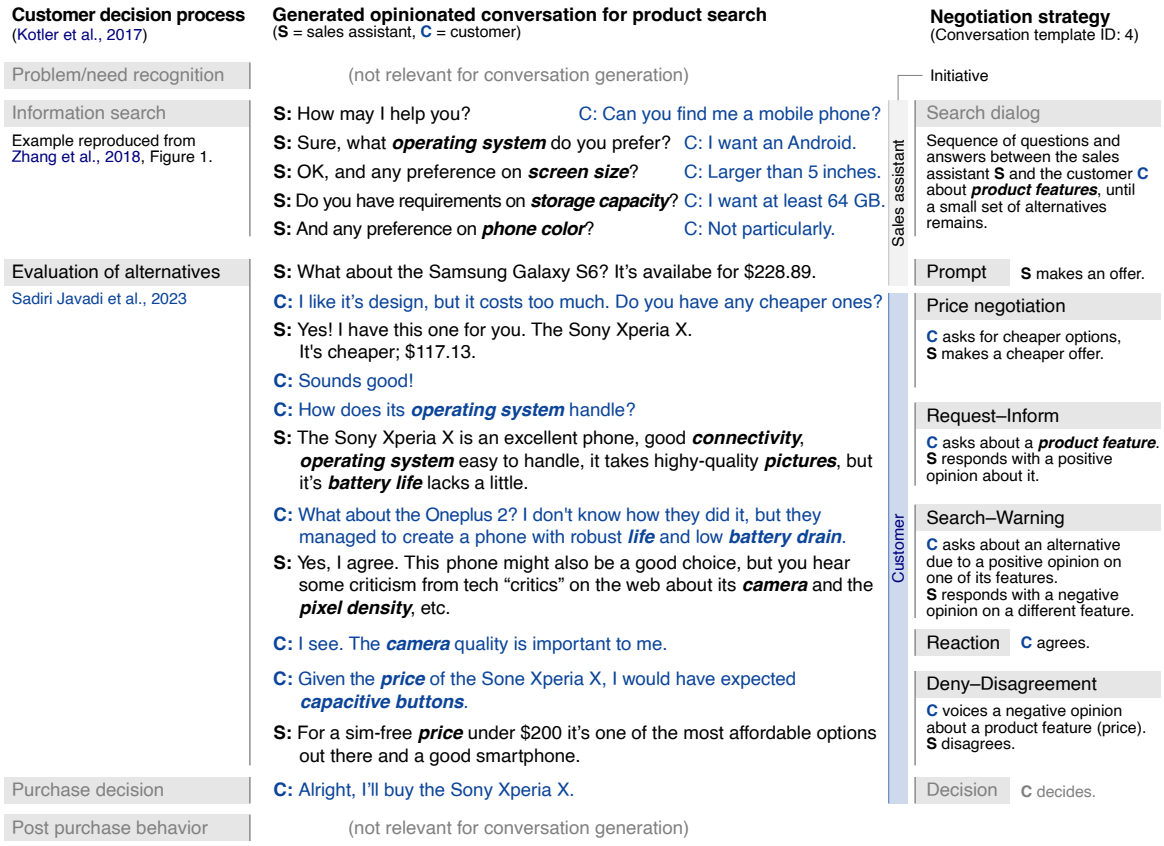

Figure 9. Illustration of the sales conversation simulation scheme by Sadiri Javadi et al. [2023].

information search. In this phase, relevant product categories are explored, and when many alternatives exist, a shortlist is formed based on user preferences. This stage is largely fact-based and relatively straightforward. Third, the shortlisted alternatives are evaluated in more depth to select a specific product, followed, fourth, by the purchase decision itself. Finally, in the fifth stage, the buyer uses the product and may engage in post-purchase behavior, such as commenting on its quality. Such feedback is frequently encouraged by shopping platforms like Amazon, where user-generated reviews play a crucial role in shaping the purchasing decisions of other consumers.

Sadiri Javadi et al. [2023] study how conversational search systems can support users in the third phase of the customer decision process, which has so far received little attention in IR. More specifically, we ask how an LLM-based sales agent can form a grounded opinion about a product it has never used, given that LLMs lack real-world experience. We find that product reviews provide an excellent source of grounded opinions, which a RAG system can retrieve and quote in a conversation with users. While users are currently accustomed to consuming product reviews in a self-service manner, much like with traditional search, both the review ecosystem and shopping platforms have become bloated to some extent. A conversational sales agent may therefore offer relief by helping users navigate this space more efficiently during their decision-making process.

This raises an additional challenge: LLMs should hold consistent opinions about products. Moreover, because the customer drives the interaction and because a conversational sales agent cannot know a customer's personal preferences in advance, the agent requires a negotiation strat-

egy to serve the customer effectively. Product sales, especially for higher-priced items, involve a complex set of objectives. For a human sales agent, closing a sale is important, but not at the risk of the customer returning the product the next day. Instead, the goal is for the customer to leave satisfied with their decision, while selling the product (with the highest price) that still aligns with the customer's needs. For example, when my family recently had an had an air-conditioning system installed, the sales agent convinced us against the more expensive setup we initially considered and instead recommended a comparatively minimal solution that better fit our situation.

In our study, we compiled a range of sales negotiation strategies and simulated sales conversations between a conversational sales agent and a customer. We then asked human annotators to assess the realism of these conversations, which they generally found adequate. That said, this work only scratches the surface of what is required to make LLMs and RAG systems genuinely helpful for product sales in a user-oriented manner. We expect substantial future work to be needed to understand the subtleties involved in building effective LLM-based sales agents in the sense of primarily assisting customers, rather than merely optimizing a retailer's bottom line.

In many respects, today's RAG systems already resemble sales agents. What they "sell" is information. However, current RAG systems, whether by accident or by design, are often reported to be quite sycophantic, which appears to be a rather simplistic (but maybe effective, system-oriented) way of ensuring user retention. A user-oriented RAG system that acts as an intermediary between users and primary information sources should be designed to support deliberation and enable users to consider alternative answers instead of steering them toward one single answer.

4 The Next Turn in Information Retrieval

As LLMs proliferate into virtually all social systems, the information retrieval community is working to understand this new paradigm of information access by observing changes in information behavior and by helping to shape what may become the next stable state of IR. Whether RAG systems are here to stay, or whether they will eventually be supplanted by other paradigms such as task-oriented AI agents, remains an open question.

With their integrated information seeking and retrieval model, Ingwersen and Järvelin [2005] synthesize decades of observations of human information behavior and of how people interact with a sociotechnical information systems to satisfy their information needs. A depiction of the generic model is shown in Figure 10. It illustrates how a cognitive actor interacts with an information system through one of its interfaces in order to manipulate information objects, with the interaction being conditioned by the actor's organizational, social, and cultural context. In their book, Ingwersen and Järvelin systematically instantiate the model for the five actor roles they identify (authors, indexers, designers, selectors, and seekers) and for two types of interfaces: those used by authors (e.g., gated collections such as scientific conferences) and those used by seekers (search engines). Given that the book was written at the height of Google's success in web retrieval, reference librarians do not even appear anymore as an interface for seekers.

In light of the emergence of LLMs and RAG systems, Figure 11 presents two additional instantiations of the model. On the left, authors interacting with LLMs to create new information objects, which they then self-publish via web services. Authors and LLMs are indistinguishable, joint actors, and it becomes unclear how initiative and responsibility for the resulting information

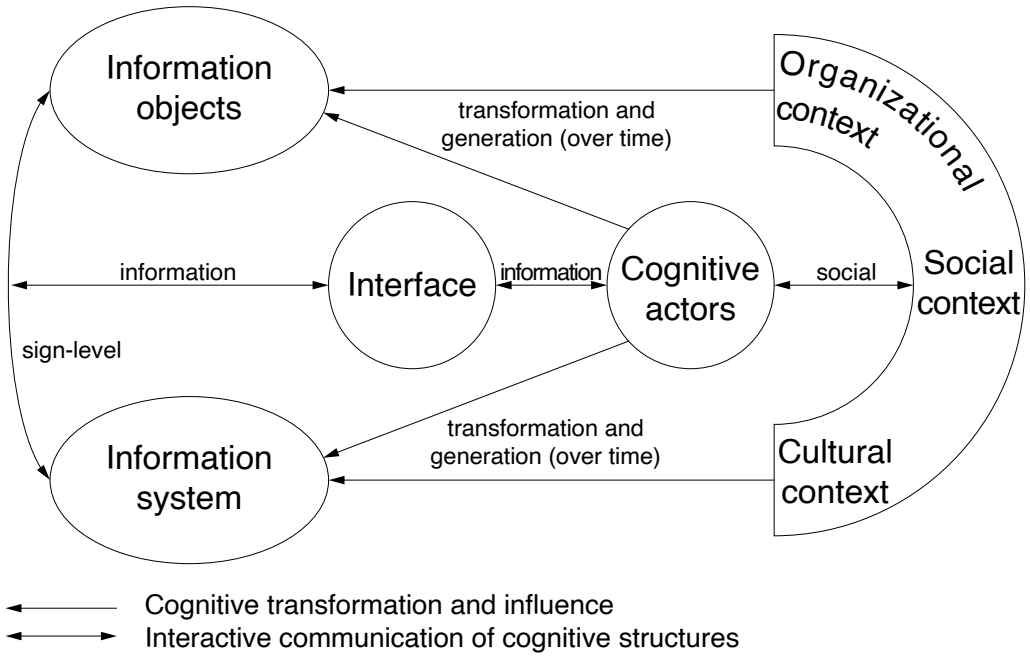

Figure 10. The information seeking and retrieval model (IS&R model) by Ingwersen and Järvelin [2005].

objects are distributed among them. In the extreme, human authors may take a back seat and merely operate automated language models to generate information objects at scale.

On the right, the role of RAG systems is depicted as that of automated reference librarians, whereas traditional search engines offer a self-service interface to the web as an information system where users act as seekers but also have to perform much of the work traditionally done by a reference librarian themselves. RAG systems shift this cognitive effort into the system. Seekers now need only articulate their information need as a prompt and are presented with an executive summary derived from a selection of information objects. In this setting, the LLM underlying a RAG system searches for relevant information objects, assesses their relevance, and synthesizes a multi-document summary that is presented to the user as a direct answer.

Five years ago, as featured snippets increasingly appeared as direct answers on search engine results pages, we formulated the Dilemma of the Direct Answer as "a user's choice between convenience and diligence when using an information retrieval system" [Potthast et al., 2020]. We expressed concern that presenting a snippet from the single most relevant web page creates a strong incentive for users to accept it at face value. In retrospect, however, this development is a natural continuation of a much longer trend. From its inception, information retrieval has sought to maximize convenience and minimize users' cognitive effort. After all, the very notion of ranking exists so that users need not browse all but only a few documents; ideally only one.

Today, RAG systems and deep research agents take over nearly all of the cognitive overhead involved in searching, browsing, and synthesizing answers. At the same time, this new interface conceals much of the underlying process and deprives seekers of developing their own sense of a

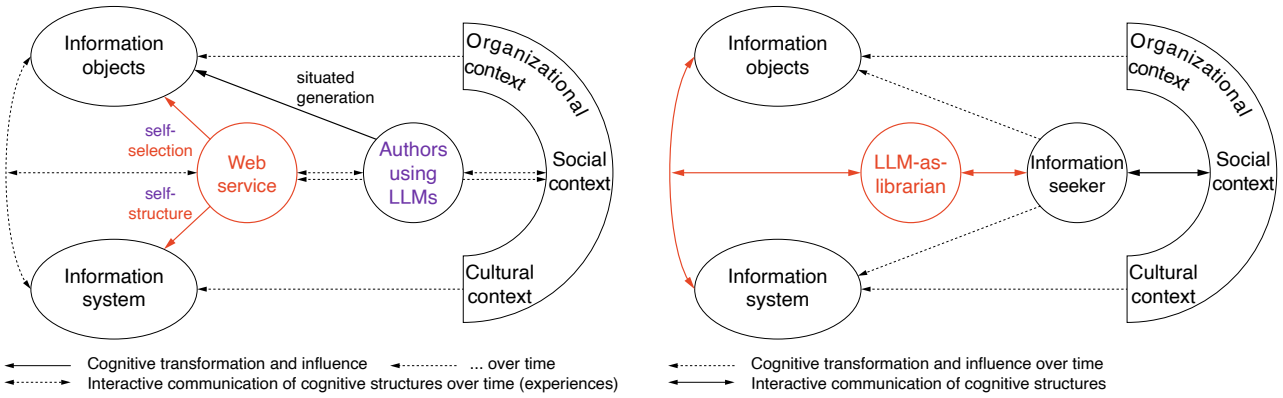

Figure 11. The IS&R model for authors using LLMs (left) and for RAG systems (right).

question's difficulty or of how validity and reliability are established. By removing the struggle of browsing results, searching may no longer function as an inherent learning exercise.

But it would be too simplistic to view human seekers using RAG systems as merely passive consumers of information, and it would be patronizing to design systems that force users to work harder "for their own good." Hardly anyone particularly enjoys using a search engine for its own sake; searching is a response to a perceived need. Moreover, the more complex the information need, and the higher the stakes of obtaining a correct answer, the more carefully users will engage, even with RAG systems. Finally, as cognitive effort is freed on the part of the seeker, that capacity can be invested elsewhere.

5 Conclusion

This brings the premise of my talk back together with the discussion of the potential and limits of LLM detection and LLM advertising. We argue that in the future diligent seekers will need to invest an increasing share of their newly freed cognitive effort into assessing the authenticity of online information and more specifically into evaluating different aspects of the authenticity of the answers generated for them. The examples I have presented are not exhaustive, and there are likely many additional facets of authenticity for which new technologies can and should be developed.

The task of information retrieval has not changed. It remains as challenging as ever: to convey through the narrow channel of human perception the information needed to fill a knowledge gap, accomplish a task, achieve a goal, or spark curiosity. What has changed is the surrounding information landscape. The technologies required to build effective information systems in this new setting will have to incorporate mechanisms for assessing and supporting information authenticity.

Acknowledgments

I would like to thank the ICTIR 2025 chairs for their kind invitation which gave me the opportunity to contemplate our recent research. I am deeply grateful to all members of the Webis Group, without whom none of this research would have been possible. I would like to thank Janek Bevendorff in particular for his long-standing dedication to pushing the limits of authorship analysis and for developing new theoretical perspectives on LLM detection. I also thank Matti Wiegmann for his diligent work together with Janek to keep PAN running at CLEF over the past years as a primary evaluation venue for authorship analysis research. Further thanks go to Janek, Matti, as well as Sebastian Heineking and Ines Zelch for their recent work on analyzing advertising and its past and future impact on search systems. Finally, I am especially thankful to Benno Stein and Matthias Hagen for their continued contributions to all of these lines of work and for sustaining the Webis Group as a whole.

References

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski, editors, *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, volume 10772 of *Lecture Notes in Computer Science*, pages 820–824, Berlin Heidelberg New York, March 2018. Springer.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. Heuristic Authorship Obfuscation. In Anna Korhonen, Lluís Màrquez, and David Traum, editors, *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1098–1108. Association for Computational Linguistics, July 2019a. URL https://aclanthology.org/P19-1104/.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Generalizing Unmasking for Short Texts. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 654–659. Association for Computational Linguistics, June 2019b. doi: 10.18653/v1/N19-1068. URL https://aclanthology.org/N19-1068/.

Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024)*, volume 14610 of *Lecture Notes in Computer Science*, pages 56–71. Springer, March 2024. doi: 10.1007/978-3-031-56063-7_4.

Janek Bevendorff, Matti Wiegmann, Emmelie Richter, Martin Potthast, and Benno Stein. The Two Paradigms of LLM Detection: Authorship Attribution vs. Authorship Verification. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) (Find-*

ings), pages 3762–3787. Association for Computational Linguistics, July 2025. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.194/.

Richard Brooks. Whodunnit? J.K. Rowling's secret life as wizard crime writer revealed. The Sunday Times, 2013. URL https://web.archive.org/web/20130720051842/http://www.thesundaytimes.co.uk/sto/news/uk_news/Arts/article1287513.ece.

Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. How people use chatgpt. NBER Working Paper w34255, National Bureau of Economic Research, September 2025. URL https://ssrn.com/abstract=5487080. Available at SSRN; accessed 22 Dec 2025.

Peter Ingwersen and Kalervo Järvelin. *The Turn – Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Kluwer International Series on Information Retrieval*. Kluwer, 2005. ISBN 978-1-4020-3850-1. doi: 10.1007/1-4020-3851-8. URL https://doi.org/10.1007/1-4020-3851-8.

Patrick Juola. The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digit. Scholarsh. Humanit.*, 30(Suppl-1):i100–i113, 2015. doi: 10.1093/LLC/FQV040. URL https://doi.org/10.1093/llc/fqv040.

Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In Carla E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi: 10.1145/1015330.1015448. URL https://doi.org/10.1145/1015330.1015448.

Philip Kotler, Kevin Lane Keller, and Marc Oliver Opresnik. *Marketing-Management*. Pearson Deutschland, 15 edition, 2017. URL https://elibrary.pearson.de/book/99.150005/9783863267742.

Nathan Lambert. Chatgpt: The agentic app. Interconnects (Substack), September 2025. URL https://www.interconnects.ai/p/the-agentic-app. Accessed 22 Dec 2025.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 393–407, Berlin Heidelberg New York, March 2016. Springer. doi: 10.1007/978-3-319-30671-1_29.

Martin Potthast, Matthias Hagen, and Benno Stein. The Dilemma of the Direct Answer. *SIGIR Forum*, 54(1), June 2020. ISSN 0163-5840. doi: 10.1145/3451964.3451978.

Vahid Sadiri Javadi, Martin Potthast, and Lucie Flek. OpinionConv: Conversational Product Search with Grounded Opinions. In David Schlangen, Svetlana Stoyanchev, Shafiq Joty, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, *24th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 66–76. Association for Computational Linguistics, September 2023. URL https://aclanthology.org/2023.sigdial-1.6.

Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Detecting Generated Native Ads in Conversational Search. In Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw, editors, *33rd Web Conference (WWW 2024)*, pages 722–725. ACM, May 2024. doi: 10.1145/3589335.3651489.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, 13 (7):1443–1471, 2001. doi: 10.1162/089976601750264965. URL https://doi.org/10.1162/089976601750264965.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *J. Assoc. Inf. Sci. Technol.*, 60(3):538–556, 2009. doi: 10.1002/ASI.21001. URL https://doi.org/10.1002/asi.21001.

tagesschau.de. KI statt Couch?, 2023. URL https://web.archive.org/web/20230928063255/https://www.tagesschau.de/wissen/forschung/ki-psychotherapie-100.html.

Wikipedia. Generative engine optimization, 2025. URL https://en.wikipedia.org/w/index.php?title=Generative_engine_optimization&oldid=1328028469.

Ines Zelch, Matthias Hagen, and Martin Potthast. Commercialized Generative AI: A Critical Study of the Feasibility and Ethics of Generating Native Advertising Using Large Language Models in Conversational Web Search. In Michael Granitzer, Christian Guetl, Christine Plote, Stefan Voigt, and Andreas Wagner, editors, *5th International Symposium on Open Search Technology (OSSYM 2023)*. International Open Search Symposium, October 2023. URL https://e-publishing.cern.ch/index.php/OSSYM.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 177–186. ACM, 2018. doi: 10.1145/3269206.3271776. URL https://doi.org/10.1145/3269206.3271776.