POLITECNICO DI TORINO
Repository ISTITUZIONALE

Information Retrieval in the Age of Generative AI: The RGB Model

Original
Information Retrieval in the Age of Generative AI: The RGB Model / Cornacchia, A., Galante, F., Garetto, M., Leonardi, E., Nordio, A., Tarable, A.:.. - ELETTRONICO. - (2025), pp. 602-612. (SIGIR '25: The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval Padova (Ita) July 13 - 18, 2025) [10.1145/3726302.3730008].

Availability:
This version is available at: 11583/2999757 since: 2026-02-11T12:39:25Z

Publisher:
ACM

Published
DOI:10.1145/3726302.3730008

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026



Information Retrieval in the Age of Generative AI: The RGB Model

Michele Garetto
University of Turin
Torino, Italy
michele.garetto@unito.it

Alessandro Cornacchia
KAUST
Thuwal, Saudi Arabia
alessandro.cornacchia@kaust.edu.sa

Franco Galante
Politecnico di Torino
Torino, Italy
franco.galante@polito.it

Emilio Leonardi
Politecnico di Torino
Torino, Italy
emilio.leonardi@polito.it

Alessandro Nordio
Consiglio Nazionale delle Ricerche
Torino, Italy
alessandro.nordio@cnr.it

Alberto Tarable
Consiglio Nazionale delle Ricerche
Torino, Italy
alberto.tarable@cnr.it

Abstract

The advent of Large Language Models (LLMs) and generative AI is fundamentally transforming information retrieval and processing on the Internet, bringing both great potential and significant concerns regarding content authenticity and reliability. This paper presents a novel quantitative approach to shed light on the complex information dynamics arising from the growing use of generative AI tools. Despite their significant impact on the digital ecosystem, these dynamics remain largely uncharted and poorly understood. We propose a stochastic model to characterize the generation, indexing, and dissemination of information in response to new topics. This scenario particularly challenges current LLMs, which often rely on real-time Retrieval-Augmented Generation (RAG) techniques to overcome their static knowledge limitations. Our findings suggest that the rapid pace of generative AI adoption, combined with increasing user reliance, can outpace human verification, escalating the risk of inaccurate information proliferation across digital resources. An in-depth analysis of Stack Exchange data confirms that high-quality answers inevitably require substantial time and human effort to emerge. This underscores the considerable risks associated with generating persuasive text in response to new questions and highlights the critical need for responsible development and deployment of future generative AI tools.

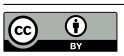
CCS Concepts

• **Information systems** → **Web searching and information discovery**; **Evaluation of retrieval results**; **Novelty in information retrieval**; • **Computing methodologies** → *Natural language generation.*

Keywords

Web answering, Information quality, Large Language Models, Retrieval-Augmented Generation, Automation bias, Stack Exchange.

ACM Reference Format:
Michele Garetto, Alessandro Cornacchia, Franco Galante, Emilio Leonardi, Alessandro Nordio, and Alberto Tarable. 2025.Information Retrieval in the Age of Generative AI: The RGB Model. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3730008



This work is licensed under a Creative Commons Attribution 4.0 International License.

SIGIR '25, Padua, Italy

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730008

1 Introduction

In recent years, the emergence of Generative Artificial Intelligence (GAI) fueled by Large Language Models (LLMs) has greatly enhanced our abilities in retrieving information and interacting with digital content. The great success and widespread adoption of AI chatbots such as ChatGPT, Copilot, Gemini, Claude, Perplexity, Llama and many others are transforming both the way we search for information and the way we produce new digital content.

The rapid growth in data availability and advances in computing power have enabled LLMs to train on huge corpora, allowing them to generate text that is not only contextually relevant but also semantically rich. These capabilities have been used to enhance search engines and virtual assistants, providing users with more intuitive and ready-to-use answers to their queries.

While companies, researchers and individuals are striving to harness the potential of LLMs, serious concerns have also arisen about the potential for misuse and unintended consequences of these tools, such as the spread of misinformation and fake news. Indeed, AI chatbots can generate responses that are syntactically and grammatically sound but factually incorrect. Furthermore, LLMs can quickly produce coherent and persuasive text that tends to be presented as the definitive answer to the users. This can increase the risk that users trust the information provided by these models without critically evaluating its accuracy or checking the original sources, thus fostering a culture of *intellectual laziness*. Also, the black-box nature of many LLM systems complicates efforts to ensure transparency and accountability in information dissemination.

Despite the vast amounts of information on which they are trained, LLMs face limitations when dealing with new or rapidly evolving topics. Since the training process relies on historical data, there is often a lag before the latest information is incorporated into the training datasets. This gap can make LLMs less reliable in answering questions about current events or emerging topics. To mitigate these limitations, some LLMs are designed to integrate with traditional search engines or databases to access up-to-date information. This approach, commonly referred to as Retrieval-Augmented Generation (RAG), allows them to gather information from the internet in real-time and provide answers to questions that fall outside the scope of their training. However, this integration is not always transparent to the user. When an LLM relies on external

search engines, it may not clearly indicate that the answer comes from a real-time search and not from its internal knowledge.

Another insidious challenge in the use of LLMs is the risk of "autophagy" in training cycles. This term, borrowed from biology, refers to a process where AI-generated content becomes part of the training data for future models, creating a self-consuming loop. As LLMs produce vast quantities of text that are indistinguishable from human-written content, this generated content often finds its way back into the public domain. Over time, the inclusion of AI-generated content in training datasets could lead to a dilution of the originality and diversity of the generated responses, as models recycle their own interpretations rather than drawing from a diverse set of authentic, human-generated sources.

The Web, as a primary source of information for billions, is especially vulnerable to the careless exploitation of emerging GAI technologies. As we continue to integrate GAI tools into information retrieval and content creation, it is imperative to balance their innovative potential with responsible oversight to safeguard the integrity of information on the Internet. It is thus crucial to develop proper methodologies and guidelines to navigate this rapidly evolving landscape and better understand its future implications.

1.1 Paper Contributions

This paper proposes a novel analytical framework to describe and understand the new dynamics of information retrieval and dissemination triggered by the integration of GAI tools into daily workflows. To the best of our knowledge, our study presents the first-of-this-kind quantitative approach to incorporate the several factors at play into a comprehensive framework capable of predicting the temporal trajectory of crucial key performance indicators.

Specifically, we focus on the generation, retrieval, and replication of answers related to novel topics for which no prior information is available — a scenario that poses the greatest challenge and risk to any answer-generating system. In this context, we examine the competition between conventional search engines and emerging generative AI systems employing hybrid strategies to generate answers, particularly real-time RAG. Our analysis incorporates both algorithmic and human behavioral aspects. The primary objective is to forecast potential future scenarios if current trends in the adoption of generative AI tools persist.

Our analytical model is complemented by an in-depth examination of a large dataset from the Stack Exchange platform, containing questions and answers in computer science and mathematics. This investigation provides important insights into the intricate temporal dynamics of answers to novel and challenging questions. In particular, it highlights the considerable time often required for high-quality answers to surface.

2 Related Work

Opportunities and risks of LLMs (more generally, foundation models) have been broadly discussed in [6, 16]. In [30] authors point out the potential for users to 'anthropomorphize' GAI tools, leading to unrealistic expectations of human-level performance on tasks where in fact these tools cannot match humans. The implications of LLMs being increasingly incorporated into scientific workflows are considered in [5]. Several studies have shown the risk for machine learning models to acquire factually incorrect knowledge during training and propagate incorrect information to generate content [3, 7, 9, 17, 25]. A framework to evaluate LLMs' ability to answer complex questions is proposed in [34]. While [29, 38, 43] explore the disinformation capabilities of LLMs, i.e., their ability to generate convincing outputs that align with dangerous disinformation narratives. The hallucination problem in natural language generation has been a focus of several empirical studies [18, 28, 36, 42]. Challenges posed by time-sensitive questions are explored in [10, 19, 21]. Techniques to detect generic LLM-generated text are presented in [27, 35], while [11, 23] focus on methods to prevent and detect misinformation generated by LLMs. A self-evaluation technique to enhance factuality is presented in [41], while [2] investigates the use of human feedback in LLM alignment.

Retrieval-Augmented Generation was first proposed in [22] for generic knowledge-intensive NLP tasks. Work [39] proposes an adaptive mechanism to control access to external knowledge by a conversational system. The use of search engine outputs to increase LLM factuality is evaluated in [37]. The problem of explaining RAG-enhanced LLM outputs has been addressed in studies such as [26] and [33], highlighting the importance of interpretability in these hybrid systems.

The risk of autophagy in model training has received significant attention, also from a theoretical point of view. The problem was first raised in the field of image processing [1, 4, 24], where feeding synthetic data back into models has been shown to decrease both the quality and diversity of generated images. Shumailov et al. considered this phenomenon for LLMs [32] (earlier version of their work in [31]), introducing the concept of *model collapse*. They show that various models, when trained with data generated by previous generations of the same model, forget the tails of the original data distribution (early collapse) and tend to collapse into a distribution that is very different from the original one, typically with much lower variance (late collapse).

Briesh et al. [8] consider three possible data sources for each generation of training data: the original dataset, fresh data, and data from any previous generation. Consistently across different combinations of these data sources, they observe a decrease in correctness in the first generation, which then increases across generations. However, this comes at the expense of diversity (especially for the model trained only with synthetic data), which collapses to zero.

Dohmatob et al. [13] consider a (linear) regression task as a proxy for more complex models and characterize the test error when models are subsequently retrained with synthetic data. They find that performance decreases with the number of retraining generations. Gerstgrasser et al. have extended this framework in [15], suggesting that data accumulation prevents model collapse for language models and deep generative models.

In [20] authors have examined the correctness, consistency, comprehensiveness, and conciseness of ChatGPT answers to 517 programming questions on Stack Overflow, showing that 52% of ChatGPT answers contain incorrect information, overlooked by users 39% of the time. In [12] authors show that activity on Stack Overflow has decreased since the advent of chat-like LLMs, as users rely more on generative AI tools, especially for coding questions.

Lastly, we mention the work of Yang et al. [40], who incorporate human behavior in self-consumption loops, similarly to us. They show that AI-generated information tends to prevail in information

filtering, whereas real human data is often suppressed, leading to a loss of information diversity.

3 System Model

We model for simplicity a system comprising one conventional search engine and one LLM integrated with it, using RAG to provide up-to-date answers to user queries. This situation is common to many GAI systems: for example, ChatGPT/Copilot are tightly integrated with Bing, and the Gemini model is incorporated into Google's search infrastructure.

3.1 The RGB Model

3.1.1 Answer model for a given topic. Consider a specific novel topic, for which no prior knowledge is available anywhere in the system at time zero. Similarly to the RGB color model, we assume that each potential answer to this topic is described by a length-3 vector, which represents a convex combination of the three primary colors (red, green and blue). We emphasize that this is just for the sake of an intuitive, simple illustration of our model. In general, we can have an arbitrary set $C = C' \cup C''$ of primary colors, associated to different initially generated answers, partitioned into a set C' of 'good' answers and a set C'' of 'bad' answers. Although the model can be generalized to an arbitrary number of primary colors, three is the minimum required to distinguish between 'bad' (low-quality, biased, fake) answers, conventionally represented by red, and two distinct 'good' answers (with potentially different qualities), associated with blue and green. This allows us to quantify both the accuracy and diversity of the produced answers, as discussed in Section 3.3. [1]

The relative popularity (or strength) of answers of a given color in the system (or one of its subsystems) at time t is determined by the number of 'coupons' of the same color in the considered subsystem.

We first consider the simpler case in which possible answers available in the system can only be red (1,0,0), green (0,1,0) or blue (0,0,1). In Subsection 3.4, we will present an extension of the model in which mixed colors (mixing some percentages of red, green, and blue) can be produced by either machine or human behavior.

3.1.2 System compartments. To specify how answers are initially generated, and how they get replicated and reinforced across different digital resources over time, we introduce five subsystems (compartments) denoted by letters G, W, T, S, L (see Figure 1). At any given time $t \geq 0$, subsystem $s \in \mathcal{S} = \{G, W, T, S, L\}$ holds $N_c^s(t)$ 'coupons' of color c, $c \in C$. We denote by $N^s(t) = \sum_c N_c^s(t)$ the total number of coupons in subsystem s, at time t. The five considered compartments are:

External sources (G): this is a virtual place representing all information sources, such as news organizations, economists, political analysts, researchers, etc., from which answers are initially generated, i.e., introduced (through human intervention) into the digital ecosystem.

Web (W): this compartment represents the set of all online resources on the World Wide Web.

[1] In case we do not care about diversity of good answers, we could implement a simpler binary system with just two colors.

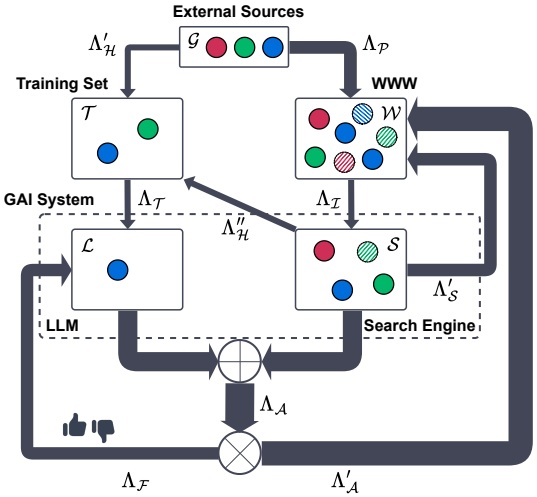


Figure 1: RGB model illustration: Colored circles represent coupons of answers to novel questions. Dashed circles are GAI-generated answers added to the Web. Gray arrows show generation/reinforcement/replication processes.

Training Set (T): this is the ensemble of all curated datasets used to train the LLM.

Search Engine (S): the search engine continuously crawls the Web, indexing and ranking all found documents. It is then able to algorithmically assess and quantify the relevance of each indexed document in relation to a user's query.

Large Language Model (L): the LLM constructs an internal representation of the information embedded within the training data, through a sophisticated architecture of interconnected parameters, or weights. Based on this representation, it produces a certain answer in response to a user's query.

Remark 1. *For all compartments, the fraction of coupons of color c is a measure of the strength of answer c compared to other answers. The* absolute *number of coupons of color c has a physical meaning only for compartments W and T, where it represents the number of distinct replicas of answer c found on the Web or in the Training Set, respectively. For the Search Engine and the LLM, only the* relative *number of coupons of color c matters, conveying a measure (weight) of how much that answer is considered relevant in the subsystem.*

3.1.3 Coupon generation and propagation. In our model, coupons accumulate in each subsystem due to generation, reinforcement, or replication processes (or flows), represented by arrows in Fig. 1. One exception is the external sources compartment, which is supposed to contain a fixed number N_c^G of coupons of color c, so that $g_c = N_c^G/N^G$ is the probability of selecting uniformly at random a coupon of color c. This allows us to account for an arbitrary distribution of the color of fresh answers that are initially introduced into the digital ecosystem.

We identify six main flows, indexed by $f \in \{\mathcal{P}, \mathcal{H}, \mathcal{I}, \mathcal{T}, \mathcal{S}, \mathcal{A}\}$. Each flow is modeled as a point process, characterized by either an aggregate rate Λ_f or a per-coupon rate λ_f, depending on the physical meaning of the flow. We will first introduce unbiased flow rates Λ_f (or λ_f). The effective rate at which coupons are generated,

reinforced, or replicated will be elucidated after the explication of our quality bias assumptions (Section 3.1.4).

- $\mathcal{P}$: this flow accounts for users posting fresh new answers on the Web, which is one of the ways answers are initially added to the digital information ecosystem. It is characterized by a (generally, time-varying) aggregate rate $\Lambda_{\mathcal{P}}(t)$.
- $\mathcal{H}$: this is the overall process by which curated answers are incorporated into the Training Set. This flow, characterized by a relatively low aggregate rate $\Lambda_{\mathcal{H}}(t)$, is divided into two subflows with rates $\Lambda'_{\mathcal{H}}(t) = \gamma\Lambda_{\mathcal{H}}(t)$ and $\Lambda''_{\mathcal{H}}(t) = (1-\gamma)\Lambda_{\mathcal{H}}(t)$, where $0 \le \gamma \le 1$ is a model parameter. The two rates correspond respectively to fresh answers generated and directly incorporated into the Training Set (by-passing the Web), and to answers coming from the Search Engine, reflecting the fact that dataset curators often use the results of traditional search engines as sources of information.
- $\mathcal{I}$: this process models how each answer posted on the Web is independently crawled and indexed by the Search Engine with per-coupon rate $\lambda_{\mathcal{I}}(t)$. Note that the aggregate rate at which documents containing answers pertaining to the considered topic are indexed is $\Lambda^{\mathcal{I}}(t) = N^W(t)\lambda_{\mathcal{I}}(t)$.
- $\mathcal{T}$: this process models how each answer contained in the Training Set is fed into the LLM, with per-coupon rate $\lambda_{\mathcal{T}}(t)$, resulting into an aggregate training rate $\Lambda_{\mathcal{T}}(t) = N^T(t)\lambda_{\mathcal{T}}(t)$ for answers related to the considered topic.
- $\mathcal{S}$: This process represents users submitting queries related to the reference topic to the conventional search engine, with aggregate rate $\Lambda_{\mathcal{S}}(t)$. Some of the answers obtained from the Search Engine are incorporated into documents posted again on the Web, creating a first recursive loop in the model. We denote the rate of answers fed back into the Web by $\Lambda'_{\mathcal{S}}(t)$.
- $\mathcal{A}$: This process represents users submitting queries related to the reference topic to the GAI system combining the LLM with the Search Engine, with aggregate rate $\Lambda_{\mathcal{A}}(t)$. Some of the answers obtained from the GAI system are fed back into the Web, at rate $\Lambda'_{\mathcal{A}}(t) = \alpha\Lambda_{\mathcal{A}}(t)$, $0 \le \alpha \le 1$, producing a second loop. Since many LLMs exploit the feedback provided by users (for example in the form of likes/dislikes) to further tune themselves, it results into another feedback loop of rate $\Lambda_{\mathcal{F}}(t) = \beta\Lambda_{\mathcal{A}}(t)$, which is relevant even if β is small, because $\Lambda_{\mathcal{A}}(t)$ is large and rapidly growing.

The hybrid strategy (RAG) employed by the GAI system to respond to queries is modeled by combining, for each color, the coupons contained in the LLM with those contained in the Search Engine: $N_c^A = N_c^L + N_c^S$. This way, the GAI system is able to respond to queries even if the LLM has not yet been trained with information related to the novel topic, provided that at least one answer is returned by the Search Engine. However, the GAI system might prefer information derived from the LLM knowledge base over that obtained from the Search Engine, when both are available. Since a strict priority rule could be potentially dangerous in this context, we consider the following soft preference mechanism: when an answer is incorporated into the LLM during the training process, we suppose that $w \ge 1$ coupons of the associated color are added to the LLM, rather than a single one. This approach permits modeling a generalized bias toward information stored in the LLM, without entirely disregarding external sources.

3.1.4 Bias to quality. We introduce, for each flow, a bias parameter towards the actual addition of coupons within the target subsystem of the flow. Such biases are introduced to take into account the effectiveness of algorithms/humans in promoting high-quality answers. We assume that each answer associated with a primary color component (three in the case of the RGB model) is characterized by an intrinsic quality q_c, normalized in such a way that $\sum_c q_c = 1$. Informally, we define the quality of each answer as the relative merit that would be attributed to it by a large number of independent experts having unlimited time to carefully evaluate and compare the answers. Intrinsic qualities are unknown to all actors in the system, both humans and machines. This fact is especially crucial for novel topics, for which none or insufficient effort has been made to evaluate proposed answers. Also, different actors have different willingness/ability to identify and promote quality among answers.

To model this, we assume that each flow $f \in \{\mathcal{P}, \mathcal{H}, \mathcal{I}, \mathcal{T}, \mathcal{S}, \mathcal{A}\}$, is characterized by a bias parameter C_f towards quality. Each time the flow is supposed to add a coupon of color c in the target subsystem s, we assume that this coupon addition occurs only with probability:

$$r_c^f = \frac{q_c + C_f}{\sum_{i \in C}(q_i + C_f)} \tag{1}$$

Note that $C_f = 0$ corresponds to the case in which coupon addition occurs proportionally to intrinsic quality. As $C_f \to \infty$, coupon addition becomes oblivious to quality, all answers being treated the same. One can also consider negative values of C_f (provided that $\min_c(q_c + C_f) \ge 0$), representing the intention to further penalize low-quality answers in favor of high-quality answers.

To summarize, at the times dictated by the point process, with aggregate rate $\Lambda_f(t)$, associated with flow f, going from subsystem s' to subsystem s, a coupon is chosen uniformly at random among all coupons currently stored in s'. Let c be the color of the chosen coupon. A coupon of color c is then added to the coupon stored in s with probability r_c^f.

3.1.5 Finite coupon time-to-live. In our model coupons do not stay forever in the subsystem in which they appear. With the exception of external sources, all compartments subject coupons to a finite duration of existence. The lifespan of each coupon in subsystem s follows an exponential distribution with a rate parameter $\mu_s > 0$, representing the inverse of the expected time-to-live. This stochastic removal process serves dual purposes: For the Web and Training Set compartments, it simulates the natural obsolescence of information and subsequent removal by document maintainers. In the Search Engine and LLM compartments, this mechanism reflects the necessity for periodic refreshment and validation for an answer to be considered still relevant.

3.2 Model Solution

Readers familiar with Markov Chains and queuing network will recognize that our system can be described as a multi-class, open system of four interconnected queues of type $\cdot/M/\infty$ associated to compartments T, W, L, S, fed by state-dependent external arrival processes of 'customers'. By construction, the giant Markov Chain representing the overall system state $\{N_c^s\}_{c,s}$ is ergodic, for any

positive values of arrival rates $\Lambda_f(t)$ ($f \in \{\mathcal{P}, \mathcal{H}\}$) and $\lambda_{f'}(t)$ ($f' \in \{\mathcal{I}, \mathcal{T}\}$). Unfortunately, due to the complex replication process of coupons in the system, our system does not allow a product-form solution for the stationary distribution of $\{N_c^s\}_{c,s}$. To solve the model, one can resort to discrete-event simulation. We implemented an ad-hoc simulator consisting of a simple C file [14].

We can also take a mean-field approach, and approximate the evolution of the mean number of coupon $n_c^s(t) = \mathbb{E}[N_c^s(t)]$ of each color in each compartment. Readers familiar with epidemic models such as the SIR model will recognize that our system is similar to compartmental models used in epidemiology. Following this approach, the system dynamics can be described by a set of ODEs (Ordinary Differential Equations). We define for convenience $n^s(t) = \sum_c n_c^s(t)$ and $q_c^s(t) = \frac{n_c^s(t)}{n^s(t)}$. We obtain the following system of coupled ODEs, which can be efficiently solved numerically.

$$\begin{aligned}
\dot{n}_c^T(t) &= \Lambda_{\mathcal{H}}\, r_c^{\mathcal{H}}[\gamma g_c + (1-\gamma) q_c^S(t)] - \mu_T\, n_c^T(t) \\
\dot{n}_c^W(t) &= \Lambda_{\mathcal{P}}\, r_c^{\mathcal{P}}\, g_c - \mu_W\, n_c^W(t) + q_c^S(t)\Lambda'_{\mathcal{S}}\, r_c^{\mathcal{S}} + q_c^A(t)\Lambda'_{\mathcal{A}}\, r_c^{\mathcal{A}} \\
\dot{n}_i^L(t) &= n_c^T(t)\, w\, \lambda_{\mathcal{T}}(t) - \mu_L\, n_c^L(t) + q_c^A(t)\Lambda_{\mathcal{F}}(t)\, r_c^{\mathcal{F}} \\
\dot{n}_c^S(t) &= n_c^W(t)\lambda_{\mathcal{I}}(t)\, r_c^{\mathcal{I}} - \mu_S\, n_c^{\mathcal{S}}(t)
\end{aligned} \tag{2}$$

One technical issue arising with this approach is that since $n^S(0) = 0$ (there are initially no coupon in the Search Engine), we have indeterminate $0/0$ forms for $q_c^S(0)$ and $q_c^A(0)$. We solved this problem assuming that the Search Engine contains, initially, a given positive number $n_b^{\mathcal{S}}(0)$ of 'black' coupons which, if chosen, do not produce any effect. Black coupons are never replenished, thus they gradually disappear from the system at rate μ_S: $\dot{n}_b^S(t) = -\mu_S\, n_b^{\mathcal{S}}(t)$. Besides solving the above problem, black coupons serve another purpose, i.e., they can model an answering machine (either the Search Engine or the GAI) preferring not to provide any answer (which happens when a black coupon is selected) in the initial phase in which none or very few colored coupons have been acquired: by setting $n_b^{\mathcal{S}}(0)$, one can can thus model more or less 'prudent' answering strategies in the case of novel topics. A clear trade-off arises here between limiting the dissemination of insufficiently consolidated answers (thereby mitigating the potential spread of misinformation), and maintaining user engagement and satisfaction.

Note that the above ODEs yield deterministic trajectories representing the 'average' system evolution. They cannot be used to assess the variability of performance metrics, especially in the crucial initial transient phase. To obtain a more comprehensive understanding of system dynamics, and to characterize the distribution of possible trajectories, we must resort to simulating multiple independent runs of the system.

3.3 Metrics

Several interesting metrics can be computed by solving our model over time. The fraction $\pi^s(t)$ of relevant answers contained in subsystem s, at time t, is:

$$\pi^s(t) = \frac{\sum_{c \in C'} N_c^s(t)}{N^s(t)}$$

Note that $\pi^S(t)$ and $\pi^A(t)$ provide the *accuracy* of individual pieces of information generated, respectively, by the Search Engine and the GAI system, at time t, pertaining to the considered topic.

Restricting to the set C' of relevant answers, one might also be interested in assessing the *diversity* degree of answers stored in subsystem s. Ideally, it would be desirable to achieve, for each relevant answer c, a fraction p_c proportional to its intrinsic quality q_c (renormalized among relevant answers):

$$p_c = \frac{q_c}{\sum_{c' \in C'} q_{c'}}, \qquad c \in C'$$

If we instead obtain, in subsystem s at time t, a fraction

$$\hat{p}_c^s = \frac{N_c^s}{\sum_{c' \in C'} N_{c'}^s}, \qquad c \in C'$$

of relevant answers c, among all relevant answers contained in s, we could quantify the *diversity* degree of s by some distance metric between the discrete distributions $\{p_c\}_c$ and $\{\hat{p}_c^s\}_c$. For example, adopting the total variation distance, we can simply define the *diversity* degree $\rho^s(t) \in [0,1]$ (the higher the better):

$$\rho^s(t) = 1 - \frac{1}{2} \sum_{c' \in C'} |p_{c'} - \hat{p}_{c'}^s(t)| \tag{3}$$

Besides the accuracy and diversity of the subsystems, we are especially interested in the following key performance indicators:

FIUA: Fraction of Irrelevant Used Answers. We call 'used' the answers added to the set of online resources on the Web, after having been suggested either by the Search Engine or by the GAI system. Note that we are not considering here answers found on the Web which have been initially generated (by flow $\mathcal{P}$), but only those produced by flows $\mathcal{S}'$ and $\mathcal{A}'$. FIUA is the fraction of irrelevant used answers found on the Web, with respect to all used answers found on the Web (at any given time t).

AIRI: AI Responsability Index. Restricting to the set of irrelevant used answers found on the Web (at any given time t), AIRI provides the fraction of them generated by the GAI system (i.e., over flow $\mathcal{A}'$), distinguishing them from those coming from the Search Engine (i.e., over flow $\mathcal{S}'$).

FRQ: Fraction of Responded Queries. Recall that our model incorporates a mechanism to simulate cautious answering strategies by initializing the Search Engine with a tunable number of black coupons: when a black answer is selected it is assumed that no usable response is returned to the user. Consequently, we can calculate, up to any given time t, the Fraction of Responded Queries (FRQ). Note that we can independently enable this mechanism for the Search Engine and/or the GAI system.

AIAI: AI Autophagy Index. Our system contains several cycles over which information about the novel topic can loop and reinforce itself. We are especially interested in quantifying the autophagy of the LLM, i.e., the amount of information (about the topic) stored in it, which has been generated by the GAI system in response to users' queries. At any time t, AIAI is the fraction of coupons found in the LLM, which have previously traversed flow $\mathcal{A}'$.

3.4 Model Extensions

The RGB model described so far can be extended to account for the generation of coupons of intermediate colors, blending primary components in varying proportions. This process mirrors a common phenomenon observed in both human cognition and machine learning, wherein new digital content – subsequently added to online resources on the Web – is obtained by synthesizing information

derived from multiple sources. This aspect of information synthesis and propagation potentially represents a critical element that warrants incorporation into the model. We do so in two different ways, distinguishing human and algorithmic effects. In both cases we assume that just two sources are used to synthesize a new content posted on the Web, but this assumption could be relaxed as well.

For the GAI system, we assume that each generated response is derived from a linear combination of the two sources. Specifically, a random fraction u of the first source is combined with $(1-u)$ of the second, for each component, where u follows a uniform distribution over the interval $[0, 1]$. This reflects the fact that the GAI system is not really able to prefer one source over the other (but note that sources to be mixed are first chosen proportionally to their relative strength/popularity). A unique randomly generated text is then returned to the user, who will use it 'as it is'.

We posit a different model for humans creating new content from the results returned by the Search Engine. Specifically, we propose that users combine a fraction ξ of the perceived 'best' source[2] with a complementary fraction $(1-\xi)$ of the alternative source. The parameter ξ is constrained to the interval $[1/2, 1]$, reflecting users' tendency to favor the perceived trusted source. This model encapsulates the cognitive process by which users evaluate and discriminate between the relative quality of multiple alternatives. This process demands significant human effort, which is driving more and more people to delegate the task to artificial systems.

4 Scenarios

Our model has a rich set of parameters, reflecting the utility of a versatile analytical framework for exploring the dynamics of a variety of possible scenarios. As a side effect, an exhaustive exploration of individual parameter impacts exceeds the scope of the present study. We therefore focus on select scenarios, reserving a more extensive sensitivity analysis for a forthcoming journal publication.

To demonstrate the insights that our model can provide, we focus on a fixed set of parameters, and adopt an evolutionary perspective, by which just a few parameters are shifted to reflect current trends in technological adoption. Specifically, we propose the following three scenarios:

pre-GAI: This scenario represents the conventional paradigm of information retrieval on the World Wide Web, before the availability of GAI services.

GAI: This scenario aims to capture present conditions, characterized by rapid yet judicious adoption of GAI tools. It reflects a transitional period where users maintain vigilant oversight of GAI outputs while still predominantly relying on results generated by traditional search engines.

post-GAI: This prospective scenario envisions a future state where information retrieval will be largely based on GAI services. It postulates a significant shift in user behavior, marked by increased confidence in and reliance on GAI-generated outputs, and a concomitant reduction of personal synthesis of information by employing search engines.

We consider an RGB model where primary components have intrinsic qualities $q_{\text{blue}} = 0.5$, $q_{\text{green}} = 0.4$, $q_{\text{red}} = 0.1$. The set of 'good' answers is $C' = \{\text{blue}, \text{green}\}$, while the red answer is bad, $C'' = \{\text{red}\}$. All primary components are equally likely to be initially generated, so $g_c = 1/3$, $\forall c$.

Table 1 reports the flow parameters chosen for the **GAI** scenario. The time unit of our system is the day, and we assume that all parameters remain constant over time[3]. So, for example, the first row in Table 1, which specifies the posting process $\mathcal{P}$ of new answers on the Web, indicates that, starting at time 0, answers are posted with rate of 1 per day, with a quality bias parameter $C_{\mathcal{P}} = 1$, and they persist online for a duration that follows an exponential distribution with a mean of 100 days.

flow	rate	bias	TTL	related
$\mathcal{P}$	$\Lambda_{\mathcal{P}} = 1$	$C_{\mathcal{P}} = 1$	$1/\mu_W = 100$	
$\mathcal{H}$	$\Lambda_{\mathcal{H}} = 0.1$	$C_{\mathcal{H}} = -0.08$	$1/\mu_T = 1000$	$\gamma = 0.5$
$\mathcal{I}$	$\lambda_{\mathcal{I}} = 0.1$	$C_{\mathcal{I}} = 0$	$\mu_S = \lambda_{\mathcal{I}}$	
$\mathcal{T}$	$\lambda_{\mathcal{T}} = 0.1$	–	$\mu_L = \lambda_{\mathcal{T}}$	$w = 3$
$\mathcal{S}'$	$\Lambda_{\mathcal{S}'} = 100$	$C_{\mathcal{S}'} = 0.1$	–	$N_b^S(0) = 10$
$\mathcal{A}'$	$\Lambda_{\mathcal{A}'} = 10$	$C_{\mathcal{A}'} = 0.1$	–	$\alpha = 0.4$
$\mathcal{F}$	$\Lambda_{\mathcal{F}} = 1$	$C_{\mathcal{F}} = 0.1$	–	$\beta = 0.04$

Table 1: Flow parameters of the GAI scenario.

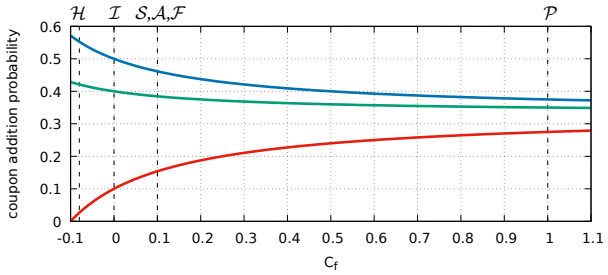


Figure 2: Impact of C_f on the coupon addition probabilities of the red, green and blue primary colors. Vertical dashed lines correspond to the values chosen for the system flows.

Before explaining the rationale behind our parameters' choice, it is useful to observe on Fig. 2 the impact of quality bias parameter C_f on the coupon addition probabilities $\{r_c^f\}_c$ of each primary component c, see eq. (1). Value -0.08, chosen for process $\mathcal{H}$, reflects the efforts of dataset curators to retain only high-quality information. Conversely, the value 1, assigned to process $\mathcal{P}$, reflects the limited attention to quality for the information initially published on the Web. Moreover, we consider that the indexing process $\mathcal{I}$ performs a rather good job at quality discrimination ($C_{\mathcal{I}} = 0$), while a slightly less effective filtering (value 0.1) is applied by users on machine-generated answers (processes $\mathcal{S}, \mathcal{A}, \mathcal{F}$).

Table 1 reveals a 10:1 ratio between rates $\Lambda_{\mathcal{S}'}$ and $\Lambda_{\mathcal{A}'}$, reflecting the assumption that, within the **GAI** scenario, users still predominantly rely on conventional search engines for verifying information incorporated into new digital content. We further notice on Table 1: i) $\gamma = 0.5$, meaning that half of the information incorporated

[2]The quality of a generic answer i with components $\{x_c^i\}_c$ is defined as the weighted average of the intrinsic qualities of its components: $q_i = \sum_c x_c^i q_c$.

[3]In actuality most parameters should be considered time-dependent due to the rapid evolution of information retrieval practices, but we adopt a simplifying assumption of constancy for the duration of the topic's lifespan, i.e., the temporal window in which the bulk of questions and answers related to the topic are generated.

into the Training Set comes from online resources; ii) $\mu_S = \lambda_I$ and $\mu_L = \lambda_T$, meaning that the strength associated to answers learned by both Search Engine and LLM is discounted over time at the same rate at which it is reinforced. iii) $\alpha = 0.4$, indicating that 40% of answers generated by GAI tools are assumed to be used in the creation of new content.

Taking this configuration as our reference **GAI** scenario, the parameters for the other two scenarios can be readily specified by difference with respect to those listed in Table 1. In the **pre-GAI** scenario, we simply set $\Lambda_{\mathcal{A}} = 0$, effectively nullifying all effects produced by users employing the GAI system.

For the **post-GAI** scenario, we envision a few fundamental parameter shifts inspired by the current trends in information retrieval practices. The primary modification involves an inversion of the rates for processes $\mathcal{S}'$ and $\mathcal{A}'$, with $\Lambda_{\mathcal{S}'}$ now set to 10 and $\Lambda_{\mathcal{A}'}$ to 100. In light of the projected dominance of GAI-based information retrieval, we establish α at 0.8 and postulate a tenfold increase in the rate of process $\mathcal{H}$ (incorporation of answers in curated datasets), setting $\Lambda_{\mathcal{H}}$ to 1. Crucially, we hypothesize that users will increasingly rely on GAI-generated responses, consequently reducing their efforts in quality discrimination when incorporating answers in new digital content. This shift is reflected in the adjustment of $C_{\mathcal{A}'}$ from 0.1 (as in the **GAI** scenario) to 1 in the **post-GAI** scenario.

We commence by presenting select findings from the fundamental RGB model, wherein color mixing is precluded; that is, neither algorithms nor users synthesize novel responses by amalgamating information from primary answers. This model is particularly suitable for queries that elicit some form of discrete, factual information not amenable to synthesis from different sources.

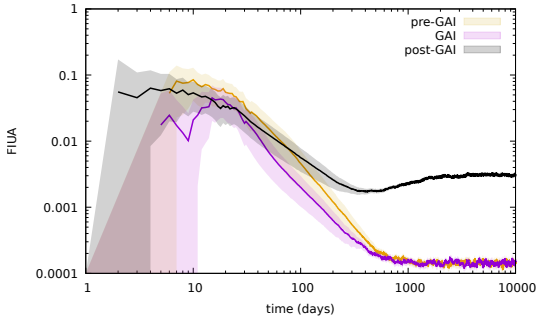


Figure 3: Temporal evolution of the FIUA indicator for the three considered scenarios. Shaded areas correspond to 95%-level confidence intervals (as in subsequent figures).

Fig. 3 illustrates, as a function of time, the fraction of irrelevant used answers (FIUA) found on the Web, under the three considered scenarios. Shaded regions surrounding each curve represent 95%-level confidence intervals derived from 400 simulation runs. These regions reveal an erratic behavior during the first ten days, after which a clear separation among the curves emerges: interestingly, the **GAI** scenario produces fewer irrelevant answers compared to the **pre-GAI** condition. However, the **post-GAI** scenario ultimately yields the least favorable outcome for this critical metric. Although the reinforcement loop of high-quality answers effectively reduces FIUA to relatively low levels across all scenarios, *the post-GAI condition generates approximately an order of magnitude more irrelevant answers than the other conditions.*

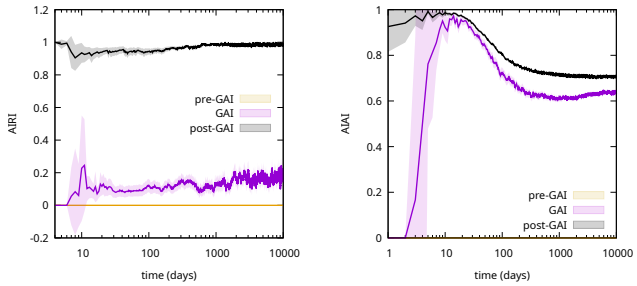


Figure 4: Temporal evolution of the AIRI (left plot) and AIAI indicator (right plot) for the three considered scenarios.

The AIRI metric (left in Fig. 4) reveals that *almost all irrelevant answers produced in the post-GAI scenario come from the generative AI system.* The AIAI indicator (right in Fig. 4) further reveals that *the LLM model is affected by a substantial degree of autophagy under both GAI and post-GAI conditions*, ranging from 60% to 70%.

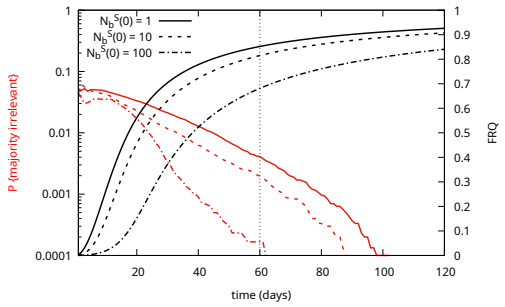


Figure 5: Probability over time that the majority of used answers on the Web are irrelevant (left y-axes), and FRQ indicator (right y-axes) in the post-GAI scenario, for initial number of black coupon in the Search Engine equal to 1,10,100.

Our system exhibits high stochasticity, particularly during the initial months of topic existence. A notable concern is the potential for irrelevant responses to gain an early, chance-driven advantage over relevant ones. To investigate this critical phenomenon, we conducted 30,000 simulations spanning the first 120 days, calculating the proportion of runs in which, at any given time t, irrelevant used answers in the Web outnumber relevant ones. Focusing on the **post-GAI** scenario, we plot this novel metric while varying the initial number $N_b^S(0)$ of black coupons in the Search Engine, which emulates more or less cautious answering strategies. Fig. 5 presents these results alongside the corresponding Fraction of Responded Queries (FRQ) metric on the secondary y-axis. A clear trade-off emerges, whereby more cautious strategies (larger $N_b^S(0)$) can significantly mitigate the probability of the aforementioned critical event, albeit at the expense of diminishing the overall fraction of responded queries (see for example the values at $t = 60$, marked by a vertical dotted line).

4.1 Results Under Answer Mixing

This section presents findings from our case study configuration, examining scenarios wherein either the GAI system or users utilizing conventional search engines synthesize novel responses by integrating two distinct information sources. This process progressively populates the system with coupons of intermediate chromatic values, containing various proportions of primary RGB components. For a detailed description of the extended model incorporating this fundamental aspect of information synthesis, readers are referred back to Sec. 3.4. Recall that, while we consider the GAI system to uniformely mix two randomly chosen answers, we instead assume that users employing the conventional search engine generate a biased combination of two random answers, applying to the best one a weight $\xi \in [1/2, 1]$. The parameter ξ thus quantifies the average discriminatory capacity of users in assessing answer quality. A value of $\xi = 1/2$ represents an unbiased mixture, indicating no discrimination, while $\xi = 1$ denotes perfect discrimination, where users consistently identify and select the superior information source.

We will separately consider the GAI and post-GAI scenarios, comparing in each scenario the effects of the following conditions:

no-mix This corresponds to the base case in which answers are never mixed. It serves as a baseline for comparative analysis.

mix-GAI We enable mixing exclusively for the GAI system. Users employing the Search Engine still use single-source responses.

mix-GAI+SE(ξ) We enable mixing for both GAI and Search Engine. We will investigate the effects of varying the quality discrimination parameter ξ, taking values of 0.5, 0.75, and 1.0.

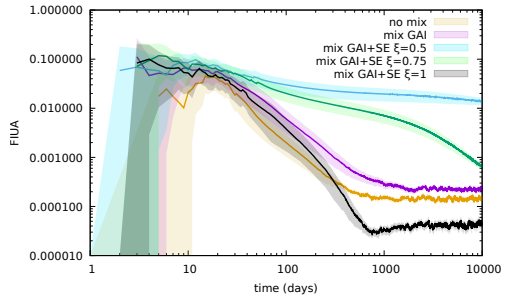


Figure 6: Temporal evolution of the FIUA indicator in the GAI scenario, under different answer mixing assumptions.

Fig. 6 reports the Fraction of Irrelevant Used Answers (FIUA) found in the WWW within the GAI scenario. Recall that here there is a 10:1 ratio between the utilization rates of answers derived from the SE compared to those generated by the GAI. We notice that the mix-GAI condition consistently yields a higher proportion of irrelevant answers compared to the no-MIX baseline. This observation suggests that the indiscriminate amalgamation of information from diverse sources generally proves detrimental to answer quality. When we also enable the mixing of SE answers by users, we observe a substantial impact of the quality discrimination parameter ξ: a markedly large FIUA is obtained when users indiscriminately combine two sources ($\xi = 0.5$). With $\xi = 0.75$, a very long time is required for the FIUA to approach levels comparable to the no-mix baseline. Perfect discrimination ($\xi = 1$) is much faster, and ultimately yields to lowest FIUA among all conditions.

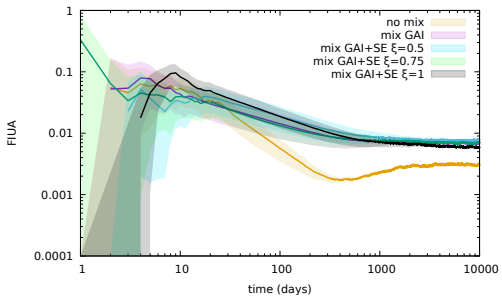


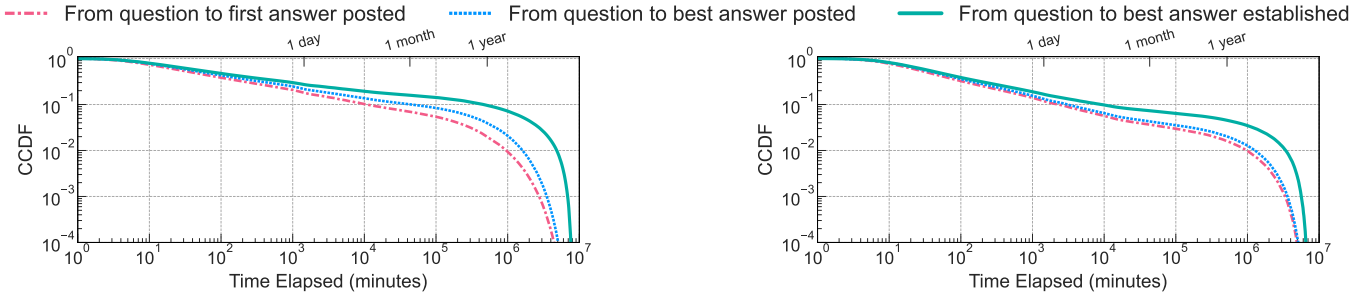
Figure 7: Temporal evolution of the FIUA indicator in the post-GAI scenario, under different answer mixing.

We conducted a similar investigation in the post-GAI scenario. Recall that here the vast majority of used answers come from the GAI system. The Fraction of Irrelevant Used Answers (FIUA), reported in Fig. 7 confirms that the generation of responses through the amalgamation of disparate sources consistently yields deleterious effects on the overall quality of information across all temporal scales. Indeed, the proportion of irrelevant information on the Web roughly doubles across all considered mixing hypotheses, due to the reduced impact of search engine-generated responses. Here the quality discrimination performed by users has marginal impact because only few users still employ results by the Search Engine.

5 Stack Exchange Analysis

Many generative AI systems opt for a chat-based interaction with the user and generally tend to provide a singular, authoritative response even to novel questions. This precludes the presentation of diverse perspectives and effectively suppresses the competition of ideas that might arise from multiple potential answers. To find out to what extent this modality can lead to suboptimal question-and-answer (Q&A) systems, we investigated one of the most well-known Q&A platforms: Stack Exchange. In particular, we looked at the largest sub-community, StackOverflow, which mainly deals with computer science and coding, and the largest (non-computer science-related) community, MathStackExchange, which covers a wide range of mathematical topics.

On Stack Exchange, users submit questions, and community members provide answers, which can then be evaluated through upvotes and downvotes. This crowdsourced rating system determines the ranking of answers, with the highest-rated response, hereafter referred to as *best* answer, displayed first, followed by others in descending order of their accumulated score. We extracted 11,555,969 questions from StackOverflow together with the corresponding 20,166,328 answers, generated by over 1 million users. Similarly, we extracted 1,068,196 MathStackExchange questions and their 1,493,849 answers. The aggregated size of the raw dataset is about 150 GB. We processed this extensive dataset on a HPC cluster, utilizing the Dask Python library to flexibly manage parallel computation. All code utilized in the data processing phase will be made publicly accessible [14].



(a) StackOverflow **(b) MathStackExchange**

Figure 8: Complementary cumulative distribution function (CCDF) of the time it takes for an answer to be posted (dash-dotted), for the best answer to be posted (dotted), and for the best answer to emerge (solid) (log-log scale).

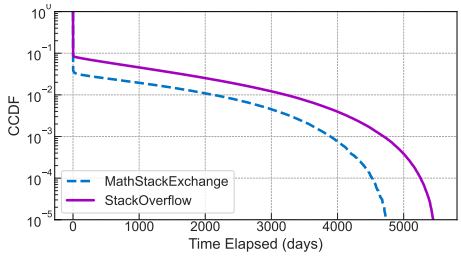


Figure 9: CCDF of the time required, in days, for the best answer to be *recognized* as superior for MathStackExchange (dashed) and StackOverflow (solid).

We found that while a large proportion of questions receive the best answer within minutes[4], a significant subset requires an extended period – sometimes spanning years – for the optimal response to emerge (for 10% of StackOverflow questions, the *best* answer is only recognized after 1 year). Figure 8 illustrates the complementary cumulative distribution function (CCDF) of three key time intervals: i) the duration until the first answer to a question is submitted (dash-dotted line); ii) the period until the best answer is posted (dotted line); iii) the interval until the best answer achieves its primacy in user display order (solid line). All intervals above are are calculated from the initial question submission time. Figure 9 complements previous results by depicting the CCDF of the amount of time required for the best answer to surface, measured relative to its initial posting time.

It is evident that there is a significant subset of questions requiring an extended temporal interval and significant collective human effort, both for the generation of high-quality responses and for the community's ultimate recognition of their superiority.

6 Discussion and Conclusions

To the best of our knowledge, ours is the first attempt to capture into a relatively simple analytical model high-level dynamics of topic-specific information across different digital resources, highlighting potential emergent phenomena and risks arising from evolving information retrieval behaviors. To this purpose, we had to strike a balance between simplicity and representational power.

Our model incorporates parameters of various natures, which are difficult to determine from real data for several reasons:

(1) Some parameters are defined at a high aggregation level (e.g. the global rate of information on a topic indexed by search engines or included in training sets). These are difficult to obtain from Q&A datasets that do not track global-scale dynamics.
(2) Some are unknown (such as the rate and accuracy at which digital resources are crawled or fed into LLMs), as private companies keep their algorithms and metrics strictly confidential.
(3) Some relate to the intrinsic quality of an answer, a metric that may be difficult to measure in the most general case.
(4) Some concern human behavior during an ongoing paradigm shift in information retrieval, for which reliable measurements are either missing or very limited.
(5) Most are changing rapidly, causing any dataset to become outdated within months of collection.

For all the above reasons, trying to fit the parameters in Table 1 using datasets makes little sense. The results produced by our model are not meaningful in absolute terms. They become interesting (in a relative sense) when we perform what-if analysis, i.e. when we vary a few crucial parameters while keeping the others fixed. We provided an example of what-if analysis in Sec. 4, where we compared two scenarios: GAI, describing answer dynamics for a hypothetical topic under current conditions, and post-GAI, describing dynamics of the same topic in an envisioned future state.

The impressive increase in user requests to AI-powered assistants and the associated proliferation of AI-generated content on the web require a critical examination of future scenarios. To address this pressing issue, we have introduced an initial quantitative framework to assess and project the impact of generative AI tools on the information ecosystem. Our preliminary results suggest a significant risk of increased misinformation proliferation, driven by two key factors: the generation of persuasive or authoritative answers to topics that are not sufficiently consolidated, and the tendency of individuals to reduce their cognitive effort in evaluating different answers and discriminating information quality. We hope our research will make a meaningful contribution to the ongoing discourse on the responsible development of AI technologies.

[4]The median time for the *best* answer to appear is 47 min on Math S.E. and 77 min on StackOverflow. The time for the *first* answer to appear is even less: 34 and 39 min.

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-Consuming Generative Models Go MAD. arXiv:2307.01850 [cs.LG] https://arxiv.org/abs/2307.01850

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL] https://arxiv.org/abs/2204.05862

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[4] Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. 2024. On the Stability of Iterative Retraining of Generative Models on their own Data. arXiv:2310.00429 [cs.LG] https://arxiv.org/abs/2310.00429

[5] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D. West, Qiong Zhang, Richard M. Shiffrin, Samuel J. Gershman, Ven Popov, Emily M. Bender, Marco Marelli, Matthew M. Botvinick, Zeynep Akata, and Eric Schulz. 2023. How should the advent of large language models affect the practice of science? arXiv:2312.03759 [cs.CL] https://arxiv.org/abs/2312.03759

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[7] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs.CL] https://arxiv.org/abs/2302.03494

[8] Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2024. Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. arXiv:2311.16822 [cs.LG] https://arxiv.org/abs/2311.16822

[9] Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1860–1874. doi:10.18653/v1/2021.acl-long.146

[10] Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. 2021. A Dataset for Answering Time-Sensitive Questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS '21, Vol. 1)*, J. Vanschoren and S. Yeung (Eds.). https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/1f0e3dad99908345f7439f8ffabdffc4-Paper-round2.pdf

[11] I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. arXiv:2307.13528 [cs.CL] https://arxiv.org/abs/2307.13528

[12] Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2023. Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow. arXiv:2307.07367 [cs.SI] https://arxiv.org/abs/2307.07367

[13] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. 2024. Model Collapse Demystified: The Case of Regression. arXiv:2402.07712 [cs.LG] https://arxiv.org/abs/2402.07712

[14] Franco Galante et al. 2025. RGB-model. https://github.com/Franco-Galante/RGB-model.

[15] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. arXiv:2404.01413 [cs.LG] https://arxiv.org/abs/2404.01413

[16] Josh Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. doi:10.48550/arXiv.2301.04246

[17] Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing The Factual Accuracy of Generated Text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 166–175. doi:10.1145/3292500.3330955

[18] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. doi:10.1145/3571730

[19] Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. TIQ: A Benchmark for Temporal Question Answering with Implicit Time Constraints. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1394–1399. doi:10.1145/3589335.3651895

[20] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 935, 17 pages. doi:10.1145/3613904.3642596

[21] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2024. REALTIME QA: what's the answer right now?. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2130, 19 pages.

[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[23] Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and Detecting Misinformation Generated by Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 3001–3004. doi:10.1145/3626772.3661377

[24] Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023. Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet. arXiv:2306.06130 [cs.AI] https://arxiv.org/abs/2306.06130

[25] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. doi:10.18653/v1/2020.acl-main.173

[26] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. arXiv:2203.11147 [cs.CL] https://arxiv.org/abs/2203.11147

[27] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1038, 13 pages.

[28] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10862–10878.

doi:10.18653/v1/2024.acl-long.585

[29] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1389–1403. doi:10.18653/v1/2023.findings-emnlp.97

[30] Murray Shanahan. 2024. Talking about Large Language Models. *Commun. ACM* 67, 2 (Jan. 2024), 68–79. doi:10.1145/3624724

[31] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv:2305.17493 [cs.LG] https://arxiv.org/abs/2305.17493

[32] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631, 8022 (July 2024), 755–759. doi:10.1038/s41586-024-07566-y

[33] Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2776–2780. doi:10.1145/3626772.3657660

[34] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family. In *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part I* (Athens, Greece). Springer-Verlag, Berlin, Heidelberg, 348–367. doi:10.1007/978-3-031-47240-4_19

[35] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The Science of Detecting LLM-Generated Text. *Commun. ACM* 67, 4 (March 2024), 50–59. doi:10.1145/3624725

[36] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. arXiv:2401.01313 [cs.CL] https://arxiv.org/abs/2401.01313

[37] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 13697–13720. doi:10.18653/v1/2024.findings-acl.813

[38] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation Capabilities of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 14830–14847. doi:10.18653/v1/2024.acl-long.793

[39] Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2024. Adaptive Retrieval-Augmented Generation for Conversational Systems. arXiv:2407.21712 [cs.CL] https://arxiv.org/abs/2407.21712

[40] Shu Yang, Muhammad Asif Ali, Lu Yu, Lijie Hu, and Di Wang. 2024. MONAL: Model Autophagy Analysis for Modeling Human-AI Interactions. arXiv:2402.11271 [cs.CL] https://arxiv.org/abs/2402.11271

[41] Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1946–1965. doi:10.18653/v1/2024.acl-long.107

[42] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219 [cs.CL] https://arxiv.org/abs/2309.01219

[43] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. doi:10.1145/3544548.3581318