



**Roskilde
University**

Designing a Neural Question-Answering System for Times of (Information) Pandemics

Graf, Johannes; Gino, Lancho; Kai, Heinrich; Möller, Frederik; Thorsten, Schoormann; Zschech, Patrick

Published in:
Information Systems Management

DOI:
[10.1080/10580530.2025.2507175](https://doi.org/10.1080/10580530.2025.2507175)

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):

Graf, J., Gino, L., Kai, H., Möller, F., Thorsten, S., & Zschech, P. (2025). Designing a Neural Question-Answering System for Times of (Information) Pandemics. *Information Systems Management, Latest articles*. <https://doi.org/10.1080/10580530.2025.2507175>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact rucforsk@kb.dk providing details, and we will remove access to the work immediately and investigate your claim.

Designing a Neural Question-Answering System for Times of (Information) Pandemics

Johannes Graf, Gino Lancho, Kai Heinrich, Frederik Möller, Thorsten Schoormann & Patrick Zschech

To cite this article: Johannes Graf, Gino Lancho, Kai Heinrich, Frederik Möller, Thorsten Schoormann & Patrick Zschech (31 May 2025): Designing a Neural Question-Answering System for Times of (Information) Pandemics, Information Systems Management, DOI: [10.1080/10580530.2025.2507175](https://doi.org/10.1080/10580530.2025.2507175)

To link to this article: <https://doi.org/10.1080/10580530.2025.2507175>



© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 31 May 2025.



[Submit your article to this journal](#)



Article views: 210



[View related articles](#)



[View Crossmark data](#)

Designing a Neural Question-Answering System for Times of (Information) Pandemics

Johannes Graf^a, Gino Lancho^b, Kai Heinrich^c, Frederik Möller^{d,e}, Thorsten Schoormann^{e,f}, and Patrick Zschech^g

^aChair of Business Information Systems, TU Dresden, Dresden, Germany; ^bFraunhofer Institute for Cognitive Systems IKS, Munich, Germany; ^cData-Driven Decision Support, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany; ^dData-Driven Enterprise, TU Braunschweig, Braunschweig, Germany; ^eFraunhofer-Institut für Software- und Systemtechnik ISST, Dortmund, Germany; ^fDepartment of People and Technology, Roskilde University, Roskilde, Denmark; ^gChair of Business Information Systems, esp. Intelligent Systems and Services, TU Dresden, Dresden, Germany

ABSTRACT

Based on Ingwersen's cognitive model of information retrieval interaction and natural language processing, this article presents (1) design knowledge in the form of requirements, principles, and features, and (2) an artifact to instantiate the design knowledge into a novel IT artifact for COVID-19 information retrieval. We conducted several evaluation episodes, encompassing technical validations and an experiment, to investigate the artifact's performance. Our work contributes to managing information and designing artifacts to handle situations of uncertainty.

KEYWORDS



Information retrieval; question answering; neural QAS; design science research; natural language processing


Introduction

A pandemic, such as the COVID-19 outbreak, comes with increased uncertainty, prompting individuals to seek health information to understand the present situation and its potential future impacts (Beisecker et al., 2022; O'Connor & Murphy, 2020; Shirish et al., 2021). Although there is a growing demand for health-related information, accessibility to scientific knowledge is relatively low because the documents are often inaccessible, written in a complex language, and hard to understand by the general public (Szmuda et al., 2020). Therefore, many rely on social media (Pennycook et al., 2020), which requires careful reflection. First, the rising availability of information often leads to the seeker's information overload (Laato et al., 2020; Mohammed et al., 2022). Second, this information is typically not filtered by professionals (Moravec et al., 2019), uses input from nonscientific sources (Pennycook et al., 2020), and is more likely to contain errors or outright false information. This is problematic, as once false information solidifies, it is nearly impossible to correct it (van der Meer & Jin, 2020). Due to the flourishing dissemination of false information in recent years (Laato et al., 2020; Nasery et al., 2023; Rocha et al., 2023; Wei et al., 2022), the WHO even introduced the term

“infodemic” to stress its danger to society (Ghebreyesus, 2020). Third, social media presents information based on an algorithm and not on the user's decisions concerning the source. Among the consequences of these issues are a decreased trust in media and fatal decisions concerning medicine (Moravec et al., 2019; Shirish et al., 2021). Accordingly, citizens should be equipped with tools to overcome these challenges (Nasery et al., 2023) as well as identify and assess reliable information (WHO, 2020).

Several solutions emerged to do so, including social media quality control (S. Wang et al., 2022), WhatsApp services (O'Connor & Murphy, 2020), and fact-checking platforms (Schuetz et al., 2021). However, managing the amount of information presumes expertise on a technical, a task, and a domain layer. Hence, socio-technical (Sarker et al., 2019) information systems (IS) are crafted for modeling corona data (Pietz et al., 2020) or designing corona dashboards (Recker, 2021). Advancements in information retrieval (IR) and natural language processing (NLP) (Janiesch et al., 2021; Young et al., 2018) have been used to create efficient question-answering systems (QAS). QASs manage, process, and explore text-based information, for instance, based on deep neural networks and pre-trained universal

CONTACT Thorsten Schoormann  tschoormann@ruc.dk  Department of People and Technology, Roskilde University, Universitetsvej 1, Roskilde DK-4000, Denmark

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/10580530.2025.2507175>

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

language models (Brown et al., 2020; Devlin et al., 2018). A QAS does not require manually crafting an advanced knowledge base (Abbasiantaeb & Momtazi, 2021), can achieve human-like performance (Rajpurkar et al., 2020), and enables interactions in natural language (Allam & Haggag, 2012). With the QAS's abilities to answer information seeker's questions in a more accessible manner, they can help face the above-mentioned issues when it comes to cognitive overload and general accuracy during information retrieval. Against this backdrop, we ask: *How to design a question-answering system capable of handling pandemic information?*

To answer this, we conducted a design science research (DSR) project to deduce design knowledge and implement a QAS. As a theoretical lens, we applied Ingwersen's cognitive model of IR interaction (Section 2). Following our research method (Section 3), we instantiate our design knowledge in the realm of COVID-19, based on an open research dataset covering a collection of more than 500,000 full-text research papers (L. L. Wang et al., 2020) (Section 4). For evaluation, we investigated the QAS's feasibility, validated the answer quality via experts, and compared the performance to standard search engines (Section 5). Finally, we elaborate on our implications (Section 6) and conclude with the article (Section 7).

Research foundations

Challenges in handling information: overload and misinformation

With the increasing demand but also availability of information, new challenges occur. As the first issue, *information overload* needs to be considered. It is a phenomenon decision-makers encounter when facing an over-abundance of information (Peabody, 1965). There are common characteristics that affect decision-makers (Roetzel, 2019): Quantity of information (Hiltz & Turoff, 1985), time pressure (Pennington & Tuttle, 2007; Schick et al., 1990), the user's processing ability (Saunders et al., 2017), as well as the quality (Burton-Jones & Straub, 2006) since information may be ambiguous (Schneider, 1987), redundant (Li, 2017), or complex (Bawden & Robinson, 2020). About 50% of people who participated in a study suffered from health information overload during COVID-19 (Mohammed et al., 2022), which is not limited to the general public (Casero-Ripolles, 2020) but also to the scientific community that struggles to keep up with the new publications (Brainard, 2020).

As a result of the growing body of information, we observe a flourishing amount of false information,

communicated on purpose (*disinformation*) or without purpose (*misinformation*) (O'Connor & Murphy, 2020). During the COVID-19 pandemic, people's information-sharing behavior has especially led to the spread of misinformation (Ghebreyesus, 2020). This contains either completely fabricated or reconfigured as well as twisted, recontextualized, or reworked content (Brennen et al., 2020). The WHO declared the reduction of false information a strategic goal (WHO, 2020) to respond to the pandemic's complexity and identify credible information. Especially on social media, people find it hard to distinguish fake from credible news (Pennycook et al., 2020). Brennen et al. (2020) showed that 20% of misinformation comes top-down but is responsible for 69% of engagement on social media. COVID-19 examples include a yoga guru and entrepreneur advertising herbal remedies as a cure against the virus (Ulmer, 2020), claims regarding the ineffectiveness of face masks (Hornik et al., 2021), and fake scientist Twitter accounts to spread misinformation about the altering of DNA when using the vaccine (Shirish et al., 2021). Against this, appropriate systems for information handling are demanded.

Question-answering systems

QAS, a research stream at the intersection of IR and NLP, answers questions posed by humans in natural language and gives responses based on a large set of information. The complex question-answering (QA) task must accommodate various dimensions, such as application, user, and question type (Hirschman & Gaizauskas, 2001; F. Zhu et al., 2021). Rather than retrieving a ranked list of documents to a keyword-based query, as most IR systems do, a QAS extracts concise answers from documents to queries in natural language (Allam & Haggag, 2012). QASs are an advancement, especially regarding large and complex unstructured documents and their information while offering a viable tool to reduce information overload (Olvera-Lobo & Gutiérrez-Artacho, 2010). Two classes of QAS can be differentiated (Janiesch et al., 2021):

- *Knowledge-based QASs* require the designer to encode formal knowledge and are usually limited to a domain (Jurafsky & Martin, 2009; Lan et al., 2020). They perform well on factoid questions but are restricted to the underlying database and logical model. The QAS designer needs in-depth knowledge of the domain to encode logic, graphs, or ontologies to form a structured QAS database (Yang et al., 2015). This may lead to a semantic gap between the designer, content creator, and user.

Also, knowledge-based QASs are considered counterintuitive for users because they often demand specific query languages, such as SPARQL (Abdi et al., 2018), and maintaining the underpinning database with new encodings is time-consuming

- *Neural QASs* are in contrast not restricted to a pre-defined knowledge base and therefore enhance flexibility (Huang et al., 2020). From an unstructured database, these systems retrieve a set of document text passages that appear relevant to the query and then extract the most likely answer(s) (Kratzwald et al., 2019). Neural QAS have improved due to advancements in deep neural network architectures, especially with the emergence of pre-trained language models, such as BERT (Devlin et al., 2018) or GPT-3 (Brown et al., 2020), that can be fine-tuned for a specific QA task. They can adapt to other databases as the underlying pre-trained model learns to comprehend language in general.

Most neural QASs have at least two components (F. Zhu et al., 2021). A *retriever module* filters the documents with IR-based methods and creates a shortlist. This is necessary because directly analyzing all documents within large databases is computationally inefficient. Subsequently, a *reader module* is responsible for text comprehension and finding the exact answer. It analyses multiple text passages and returns the top-n-ranked answers. Neural open-domain QAS have achieved human-like performance, with some models exceeding it, having over 90% accuracy (Rajpurkar et al., 2020). QASs serve, among other purposes, as chatbots for customer support (Sharma & Gupta, 2018) and conversational agents in medicine (Laranjo et al., 2018).

Developing QAS has gained great attention over the last years, also for COVID-19 (Alzubi et al., 2021; Esteva et al., 2021; Su et al., 2020). However, these approaches focus on architectural aspects, especially technical performance, while neglecting to derive generalized design knowledge from a socio-technical viewpoint. Such knowledge is important because the development of QAs is complex (Zschech et al., 2020), requiring expertise beyond the dominant technical ones, including task-/domain-related knowledge.

Cognitive model of information retrieval interaction as a theoretical lens

IR is a knowledge management process concerned with representing, storing, organizing, searching, and finding knowledge (Alavi & Leidner, 2001). It is characterized by the interplay between different actors, technical

components, and information objects. Employing the cognitive model of IR interactions from Ingwersen (1996) allows us to better understand how we should design our QAS. The model represents “the current user’s information need, problem and knowledge states and domain work task or interest in the form of contextual structures of causality” (p. 3). IR responds to challenges of information relevance (e.g., IR systems improve the relevance of results (D. Zhu et al., 2023)), data with huge volume and variety (Ghasemaghahi, 2017), and uncertainty (Ingwersen, 1996). Given that the model considers the socio-technical nature of IR including actors, technology, and tasks, it has already been used to design digital artifacts (Seidel et al., 2017; Sturm & Sunyaev, 2019) and fits our endeavor.

IR interactions comprise *information objects* (full texts and semantic entities), *technical artifacts* (IR system components and interfaces), and *communication* (individual users and social environment) on a cognitive level. The focus lies on users conducting IR or, as in our case, the QA task. The users’ cognitive models are among others expressed by their individual goals, information needs, and information behavior. By contrast, the cognitive models of information objects and technical artifacts are explications of the creators’ cognitive models (i.e., system designers or authors). Ingwersen (1996) argued that a fit between the different actors’ cognitive models is essential to ensure effective IR interactions, which are closely related to the task-technology fit theory (Goodhue & Thompson, 1995) or the cognitive fit theory (Vessey, 1991).

Research method

DSR is concerned with the design of socio-technical artifacts (Gregor & Hevner, 2013). Software instances represent artifacts that aim to solve a problem by building upon knowledge from theory and practice (March & Smith, 1995). Our study employed theoretical work to inform the artifact design as well as generalized insights from building the artifact (Meth et al., 2015). We adopted the DSR method from Kuechler and Vaishnavi (2008) in combination with Baustein’s detailed activities (Schoormann et al., 2024), which allows both designing artifacts and generalizing knowledge (see Figure 1). In this article, we differentiate between three forms of design knowledge: Design requirements (DR) represent the overall goals derived from theory or empiricism. Design principles (DP) are among the most frequently used mechanisms (Möller, Hansen, et al., 2022) to address these requirements by codifying prescriptive statements. Design features (DF) guide how to operationalize DPs (Meth et al., 2015) and

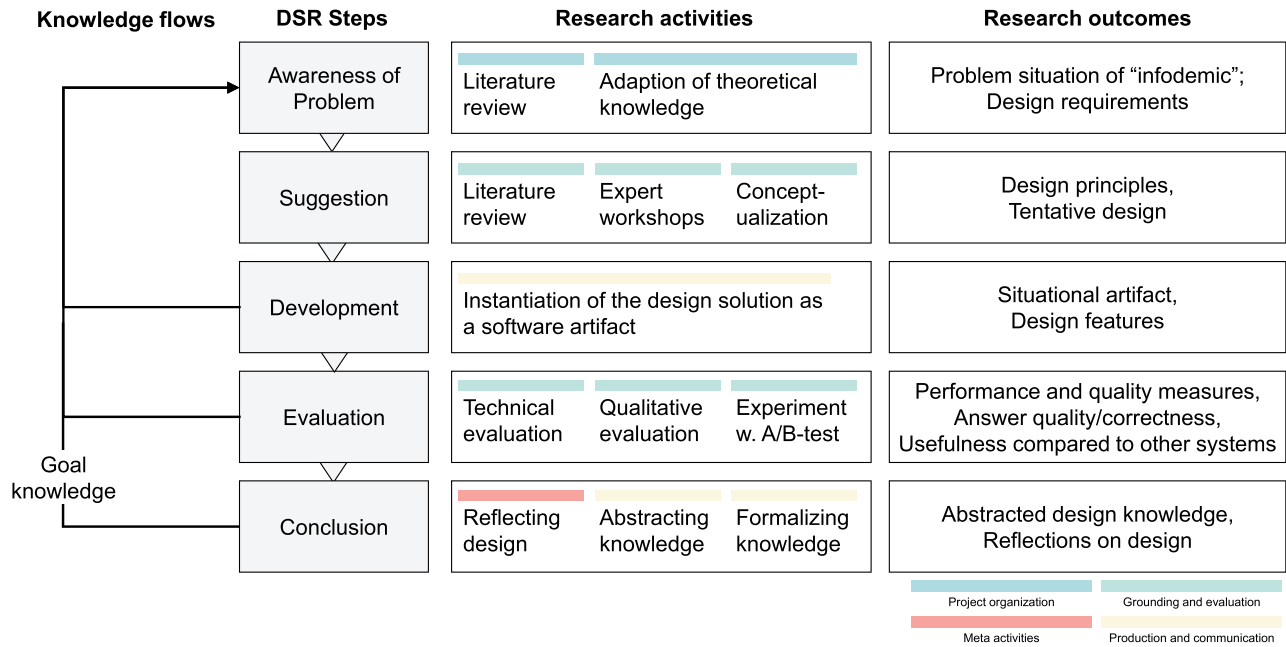


Figure 1. Overall research design.

bridge the gap between abstract knowledge and situational implementations.

Understanding the *problem situation* is a prerequisite in DSR. Based on observations, literature, and talks with experts, we recognized that the coronavirus crisis was characterized by numerous individual information needs (questions) and an abundance of generated information (answers) spread over multiple sources, including social media, news, and science. Consequently, we conceptualized the infodemic, i.e., “too much information including false or misleading information in digital and physical environments” (WHO, 2020), as a major societal problem. Guided by this problem, we employed the theoretical lens of the cognitive model of IR interaction (Ingwersen, 1996) to formulate DRs.

Based on our understanding, we *suggest* DPs to address the DRs. As QASs have become more efficient and accessible for various users, our solution is an artifact from this class as a sub-type of IR systems. To create a solid grounding, we build upon knowledge from literature and empiricism. First, we reviewed a diverse corpus of research articles. From a socio-technical view, we explored DSR studies dealing with the design of QAS/IR systems (John et al., 2016, 2018; Robles-Flores & Roussinov, 2012; Sturm & Sunyaev, 2019; Zschech et al., 2020), identified via *AISel* and *Science Direct* using the corresponding keywords in combination with “design science.” From a technical perspective, we screened academic search engines, such as *Google Scholar* and *Semantic Scholar*, for corona

QAS. Since the first technical solutions have only recently been released, we broadened our scope to include QAS surveys and conceptual papers too (Abbasiantaeb & Momtazi, 2021; Breja & Jain, 2022; Da Silva et al., 2020; Huang et al., 2020). Second, we held three workshops with a DSR scholar/NLP developer, a medical Ph.D.-student, and an undergraduate student of an economic faculty to develop our phenomenon understanding further. Participants were selected to capture different viewpoints: Information consumer vs. provider, user vs. developer, technical vs. non-technical background, and medical expert vs. medical novice. The workshops took place between October 2020 and February 2021 and lasted between 1 and 2 hours each. We were interested in the different facets of IR interactions related to COVID-19 QA. On this basis, we could advance our insights concerning diverse search behaviors in pandemic crises (e.g., ad-hoc vs. task-driven), individual information needs (yes/no vs. factoid questions), and the usage of commonly used IR systems (e.g., social media, Google Search, and PubMed).

In the *development* step, DPs were translated into more concrete DFs that represent the specific capabilities of an artifact. We report the DFs closely with the instantiation as this allows us to illustrate a specific way of operationalizing them. The instantiation serves then as the subject for the *evaluation*. To consider the socio-technical nature of our artifact, four different evaluation episodes were performed. First, computational

experiments to ensure technical feasibility (05/2021). Second, a qualitative investigation of the artifact's response quality (10/2021–11/2021). Third, an evaluation of the artifact's ability to adapt its knowledge base (e.g., consider new insights) (11/2021 and 05/2022). Fourth, an experiment with over 100 participants to determine the artifact's usefulness compared to the de facto systems Google and PubMed (06/2022–07/2022). As the evaluation results were promising, we formalized and reflected upon the knowledge gathered from the artifact's building and evaluation.

Artifact description

Conceptualization of design requirements

(Kernel) theories can be used to rigorously ground the artifact in existing knowledge and derive a set of requirements (Möller, Schoormann, et al., 2022). We conceptualized four DRs by building upon Ingwersen's cognitive model of IR interaction to respond to the infodemic.

An essential IR aspect is the cognitive load of users who search for specific information objects represented as text, pictures, etc. (Ingwersen, 1996). Users were confronted with large volumes of information from various sources in the pandemic era. For researchers, keeping up with the current level of knowledge, drawing relevant insights, and making decisions have become difficult (Brainard, 2020) since corona-related papers are published weekly within databases, such as CORD-19 (L. L. Wang et al., 2020). The general public's overload stems from the extensive media coverage (Casero-Ripolles, 2020). Given this health information overload (Mohammed et al., 2022), an artifact needs to reduce individuals' cognitive effort to fulfill their information needs. In consequence: **DR1 (Information quantity)** – *Decrease the cognitive effort of the consumer resulting from too many sources when searching for information.*

Both relevance and data veracity (Ingwersen, 1996; D. Zhu et al., 2023) need to be considered to provide useful results for a specific question. This is in line with the WHO's strategic objectives to reduce false information (Ghebreyesus, 2020). Pennycook et al. (2020) found in a study with 1,700 adults that people share false information about the pandemic because they do not sufficiently think about whether the content and the source are accurate. Hence: **DR2 (Information quality)** – *Increase the answer quality of the system to overcome the potential for intentional or unintentional spreading of false information.*

The necessary knowledge prerequisites constitute a core aspect within the cognitive model of IR

interactions as they can lead to a large gap when searching for information about an unfamiliar topic. This results in tensions, such as information from social media being accessible but often more likely subject to misinformation (Pennycook et al., 2020) vs. information from scientific sources being trustworthy but not very accessible to the public (Norman & Skinner, 2006). About 47% of EU citizens have insufficient health literacy (Sørensen et al., 2015), 9/10 adults in the U.S. struggle with health literacy (NLM, 2024), and nearly half of the adult Canadians have literacy skills below the high school level (Shahid et al., 2022). Consequently, individuals frequently lack the capability to effectively obtain, process, and understand health information, which facilitates the spread of misinformation (Cuan-Baltazar et al., 2020). Likewise, even skilled medical experts may struggle with existing tools that require a higher degree of computer literacy, for example, understanding IR techniques that work with complex operators for search queries (Vanopstal et al., 2013). Thus, a system should reduce the complexity of obtaining information objects from the medical domain to make scientific content more approachable. Accordingly: **DR3 (Information accessibility)** – *Minimize the required knowledge prerequisites and technical skills required for using the system.*

Lastly, the dynamics of information demands should be reflected. Scholars from multiple disciplines contribute to a knowledge base due to the far-reaching impact of a pandemic and the need to face new uncertainties. To provide adequate responses, it is necessary to consider numerous sub-fields, such as virology or public health relations. Montani et al. (2021) exemplified the importance of multi-disciplinary approaches to treat the long-term implications of SARS-CoV-2 patients, which is in line with prior IR literature stressing multiple sources and perspectives (Ingwersen, 1996; Seidel et al., 2017). Also, the current state of the art can change rapidly like through new vaccines or medical treatments. Therefore: **DR4 (Information variety and evolution)** – *Ensure that the system always uses the most up-to-date data sources synthesized from various research disciplines to produce reliable and up-to-date results.*

Development of design principles

New solution ideas can be obtained deductively from a kernel theory or abductively (Siering et al., 2021). We started with developing an initial set of DPs and iteratively refined them by reflecting on our design activities in conjunction with our knowledge acquisition process (theory, literature, and workshops). In the following, we introduce four DPs and provide a rationale for their

formulation. Each DP begins with characterizing the user and illustrates the context (pandemics), the mechanism, and a rationale (Gregor et al., 2020).

DP1 – interact with pandemic information in natural language

To enable users with a variety of backgrounds, such as professionals or nonprofessionals and domain experts or novices, in scenarios of pandemics to interact with the system with minimal cognitive effort, ensure that the system employs natural language search interfaces.

Rationale: Compared to IR systems in other domains, a QAS providing pandemic-related information covers many user types with varying knowledge, information needs, and information-seeking behavior. A system should thus reduce the user's cognitive effort (DR1) and minimize knowledge prerequisites (DR3). From a professional user viewpoint (e.g., medical experts), information needs are typically well-defined, mitigating uncertainty and allowing conscious navigation (Ingwersen, 1996). Contrarily, as the pandemic polarized the mainstream media, nonprofessionals were attracted (Casero-Ripolles, 2020) whose information needs are comparatively ill-defined, leading to unconscious search behavior. This requires a simple entry point to help formulate (implicit) information needs (Ingwersen, 1992; Wilson, 1999). Besides, there is a spectrum between those extreme poles in real life. The user variety implies a gap in prerequisites, vocabulary (e.g., medical terms), and syntactical structures (e.g., query language) as to why a system must map questions and answers to a shared semantic space. Providing a QAS with the ability to process natural language and a persuasive interface (Mayer & Moreno, 2003; Tawfik et al., 2014) mitigate this cognitive “free fall” (Ingwersen, 1996) to the symbolic level while retaining fractions of meaning from various input types.

DP2 – draw on reliable pandemic information sources

To enable users with varying abilities to assess the validity of crisis information (e.g., scientific vs. nonscientific), provide the system with the ability to output justifiable information objects from reliable sources so that users can rely on the quality of the output.

Rationale: Trustworthiness needs to be given (WHO, 2020) to respond to the demand for high-quality answers (DR2). Following Pennycook et al. (2020), people share false information as they do not sufficiently think about whether the sources are trustworthy. Using scientific databases can establish a credible ground truth because their reliability is shown, for instance, by the responsible curator institutions (e.g., the National Library of Medicine). Answers can be extracted from

such collections that reflect the information providers' (researchers and authors) original message and intention without distorting the content through additional information brokers. The format of the corresponding information objects (scientific papers) allows us to consider and forward the situation of a specific answer. Providing this is decisive since an answer's quality is dependent on subjective justification (Hirschman & Gaizauskas, 2001). The system should therefore provide justifiable information with the answer, allowing for a proper assessment based on its origin. This supports IR tasks concerned with judging the appropriateness of the information to the user's needs (Ingwersen, 1992; Wilson, 1999).

DP3 – represent pandemic answers with contextual details

To enable users with varying abilities to assess the quality of the answers in a specific context, provide the QAS with features to extract concise but context-embedded representations based on multiple, justified data sources so that users can estimate the answer quality.

Rationale: A user must be able to place obtained answers into a certain context (e.g., a problem situation). Thus, the system should enrich an answer with further contexts, such as insights about the paragraphs and documents from which the data were extracted. Our workshops also revealed that considering multiple answers from different documents allows us to assess answer quality better, known from data source triangulation. In this regard, the system can improve the user's confidence in the answer when it is confirmed through multiple sources. Also, providing contextual elaboration can help to debunk beliefs formed on misinformation and even stimulate protective actions in pandemics (van der Meer & Jin, 2020). Nonetheless, there is a tension between obtaining a precise answer to limit the information overload (DR1) and obtaining sufficient information to guarantee answer quality (DR2). As a result, implementing DP3 demands careful consideration.

DP4 – construct pandemic knowledge base beyond disciplinary boundaries

To enable users to consume up-to-date and cross-disciplinary pandemic knowledge, equip the QAS with backend features for flexible knowledge base construction so that users acquire synthesized knowledge agnostic of any discipline.

Rationale: The scientific community's knowledge base is steadily evolving, and researchers from many disciplines publish their work in various formats (DR4). An IR system transforms data into knowledge (Ackoff, 1989) and constructs a knowledge

representation that reflects the current state of the art. Regularly updating the provided information can reveal the scientific discourse regarding this information and strengthen the user's skepticism toward other sources of health news, decreasing the spread of misinformation (Laato et al., 2020). In this regard, the system should be able to adapt to changes in the information environment. The frequency of newly published information regarding pandemics renders systems that manually need to encode logic or ontologies to construct the knowledge base infeasible (Choi et al., 2016). DP4 emphasizes the need to incorporate learning-based retrieval and QA technologies to construct an up-to-date knowledge base dynamically.

Situational instantiation and design features

Since DPs are typically free from technical aspects, design features for bridging the gap between principles and implementations were developed (Meth et al., 2015). Next, we describe the features of each of the four DPs and our QAS prototype. Generally, the system's architecture is composed of a frontend and a backend (see Figure 2). The frontend's model-view-viewmodel (Gamma et al., 2016) supports the user's IR interaction via a graphical UI. The backend implements the application and processing logic. A REST interface middleware handles the communication between both components.

Referring to DP1, users can interact with the QAS via two features. First, selecting from ready-made frequently asked questions (FAQ) suggested by the WHO to respect rather ill-defined needs. Second, formulating free text questions to aid expert users with well-defined information needs. These features help to minimize

knowledge prerequisites by proposing (pre-defined) questions as well as accepting requests in natural language to reduce semiotic gaps and cognitive effort (see Figure 3).

After receiving user input, following DP4, the backend processes the request. The backend consists of a retriever and a reader module as part of a learning-based neural QAS (see Research Foundations). It enables flexible knowledge base construction. A Haystack pipeline (Deepset, 2021a, 2021b) served as the overall QAS framework. We implemented the retriever through a sparse Elasticsearch retriever (Deepset, 2021c) that is concerned with document relevance. The reader employs BERT (Devlin et al., 2018); a pre-trained language model allowing one to grasp general semantics and terminology of the English language that ensures scalability to various contexts. The BERT language model deployed in our instantiation was pre-trained and distilled by Turc et al. (2019) and fine-tuned by Mezzetti (2021a, 2021b) to adapt to the medical domain and increase performance on the QA task. To ensure our QAS's multi-sourcing capability, a pre-processing module for a unified document representation was implemented via JSON, treating different sources equally.

For DP2, the underpinning database ingests a sample of 300,000 documents as part of the scientific CORD-19 dataset that contains a collection of over 500,000 papers regarding coronaviruses (L. L. Wang et al., 2020). Well-known organizations, such as the National Library of Medicine, maintain the database to ensure credible ground truth data from multiple sources. We report all relevant metadata (e.g., outlet, author, institution, and publication date) from the retrieved documents,

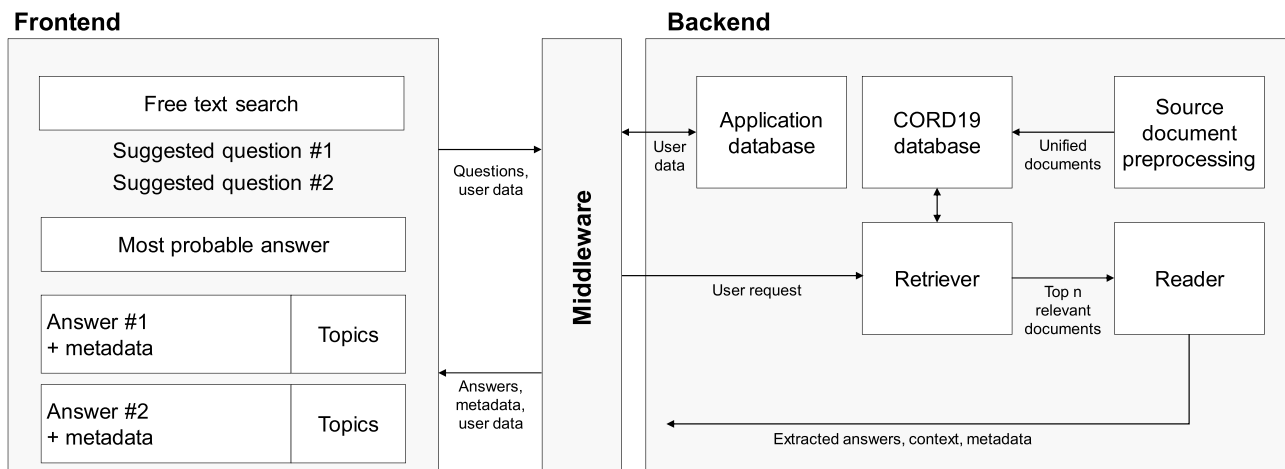


Figure 2. The QAS-architecture.

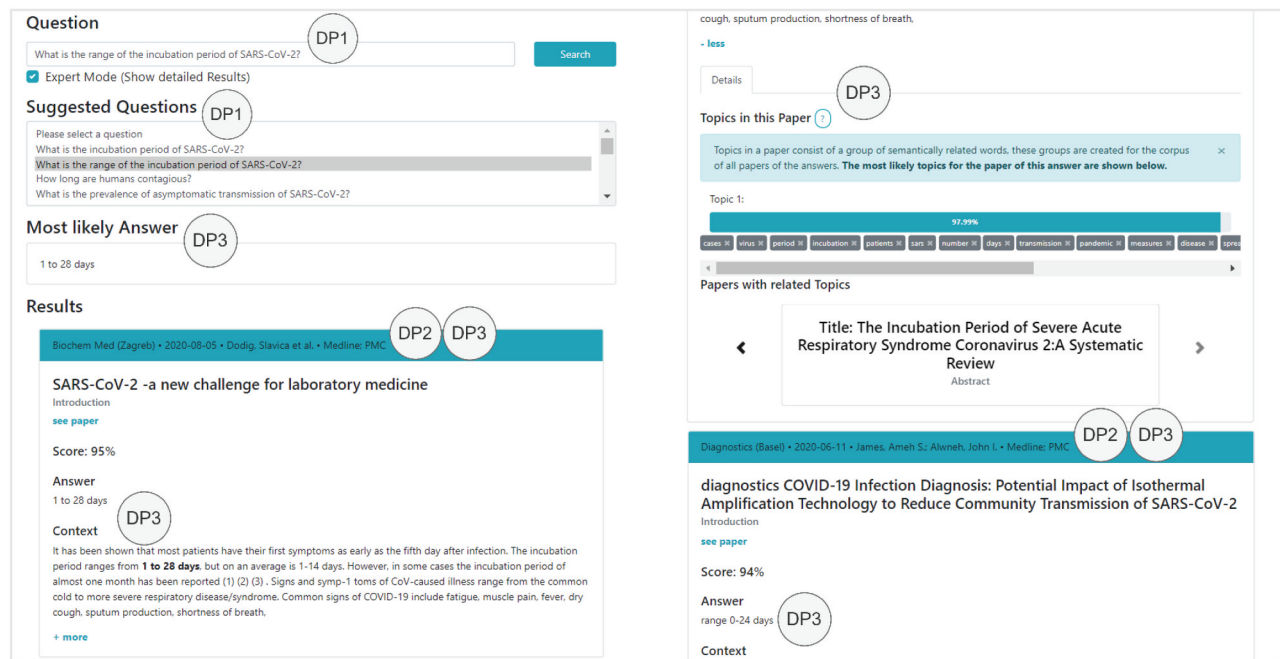


Figure 3. Search/top1 answer with context (left), topic model/top2 answer (right).

allowing us to judge the answer's origin, quality, and relevance.

When the query is processed, the user receives the response. In line with DP3, our QAS displays answers in multiple levels of granularity to ensure a concise but context-enriched representation. The core output is the presentation of the most likely answer. Further, it is possible to show the top-k answers (ranked by their probability scores) from additional documents/sources to verify their quality. The system also outlines the surrounding context (the applied language model realizes sentence-level embedding) and additional context from the meta-data. Beyond that, we implemented topic modeling via a Latent Dirichlet Allocation (Blei et al., 2003) that we trained with answers from our system. The topic model enables the user to further grasp the context of the answer by summarizing the central theme of a corresponding document with meaningful keywords. This feature seeks to reduce the cognitive effort of the user when assessing the relevance of a scientific paper.

Mapping design requirements, design principles, and design features

To summarize the obtained design knowledge, we visualized the mapping between design requirements, principles, and features in Figure 4.

Demonstration and evaluation

Considering the socio-technical nature of our artifact, we performed four evaluation episodes (Venable et al., 2016), including computational and naturalistic experiments (see Figure 5).

Episode 1: technical evaluation

As a crucial step for computational projects (Rai, 2017), we began with a technical assessment by evaluating the QAS with a COVID-QA dataset containing 2,019 question/answer pairs annotated by 15 biomedical experts (T. Möller et al., 2020). The annotations stem from 147 articles of the CORD-19 dataset and use a format similar to the widely known SQuAD2.0 dataset (Rajpurkar et al., 2018). However, this QA dataset differs in having a greater document length (avg. 6,118.5 vs. 152.2 tokens) and longer answers (avg. 13.9 vs. 3.2 words).

Our technical experiment covered the evaluation and selection of the *retriever* and the *reader* modules. We tested alternatives and measured their performance. Referring to the retriever, we compared two dense retrievers and one sparse retriever, using the IR performance metrics recall, mean average precision, mean reciprocal rank, and average retrieve time per document (Manning et al., 2008). The sparse Elasticsearch retriever achieved solid results and significantly outperformed both dense retrievers, which is why we integrated this model. For the reader, we used

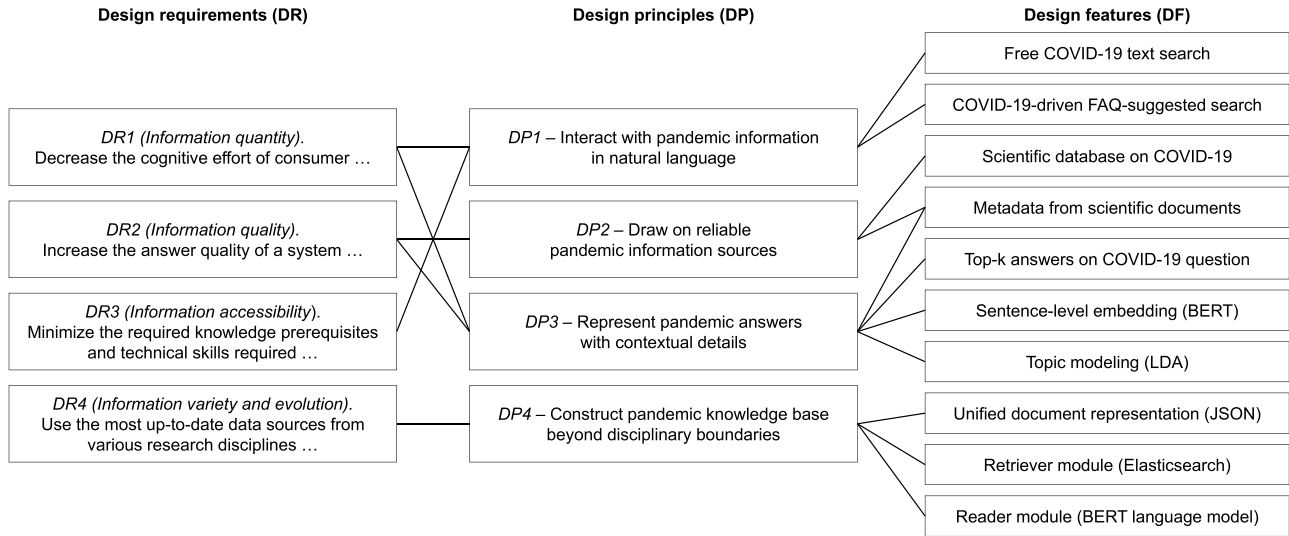


Figure 4. Design knowledge mapping diagram.

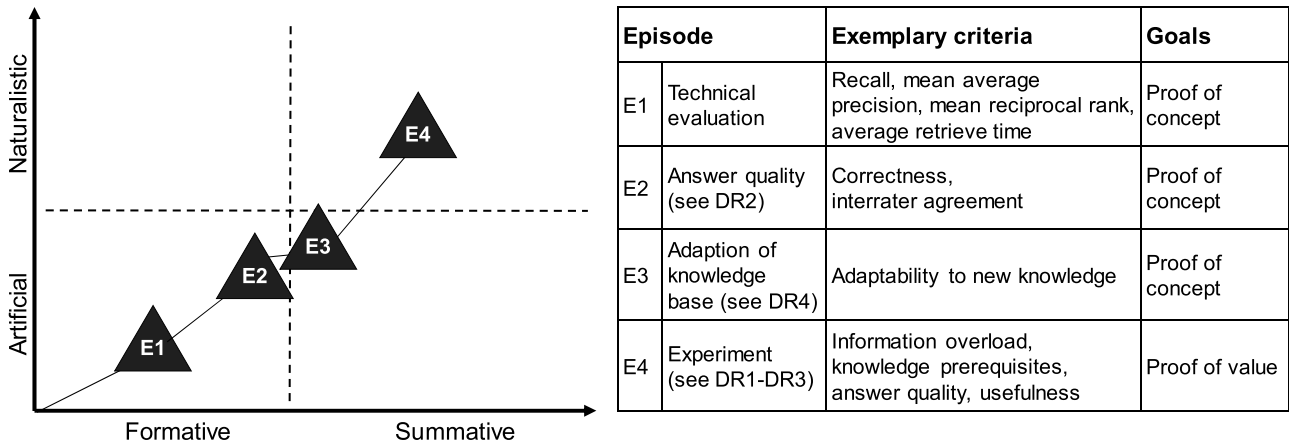


Figure 5. Overview of evaluation episodes.

a distilled version of BERT (Turc et al., 2019) with two different fine-tuning strategies and compared the results through their accuracy, exact match, F1-score, and total time required to extract answers. As a result, we implemented the version that achieved higher performance values. In Appendix A, we provide further details on the technical evaluation.

Episode 2: assessment of the answer quality

In a second episode, the quality of the answers was tested (see DR2). For this purpose, two authors of this article independently scored the correctness of the top $k = 10$ answers of our QAS for 25 questions of varying complexity, resulting in 250 question/answer pairs. The questions with corresponding ground truth (GT) answers are taken from the labeled COVID-QA dataset (T. Möller et al., 2020). We defined an answer's

correctness on four different levels. An *exact match* is semantically identical to the GT and encompasses all answer dimensions. Exact matches do not necessarily require 100% syntactical overlap. A *partial match* includes a subset of the ground truth (e.g., either “Wuhan” or “China” but not both). A *non-GT match* is a factually correct statement but differs from the GT within the COVID-QA answers. An answer is *false* when none of the previous classes apply (see Table 1).

Based on the categories, two authors classified all 250 answers. After a first run with an inter-rater agreement of 69.6%, a second run was conducted to discuss and resolve differences. After this, an agreement of 87.2% with 32 open disagreements due to a lack of medical knowledge was reached. To resolve the remaining issues, we consulted a third (external) judge with a medical background. Using the final assessment, we calculated cumulated percentages of correct answers for

Table 1. Answer categories with examples (*database version 05/2022).

Answer category	Question	Ground truth answer*	System answer*
Exact match (i)	What kind of test can diagnose COVID-19?	rRT-PCR test	PCR-based tests
Partial match (ii)	Where did SARS-CoV-2 originate?	Wuhan City, China	China
Non-GT match (iii)	Where did SARS-CoV-2 originate?	Wuhan City, China	Primary host bats
False answer (iv)	What kind of masks are recommended to protect healthcare workers from COVID-19 exposure?	N95 mask	Surgical masks

Table 2. Answer quality assessment.

Metric	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Exact match (i)	16%	36%	52%	64%	72%	72%	72%	72%	72%	72%
Partial match (ii)	44%	64%	80%	80%	88%	88%	88%	92%	92%	92%
Non-GT match (iii)	68%	84%	92%	92%	100%	100%	100%	100%	100%	100%

each category given the top k answers (see Table 2). For example, considering the top $k = 3$ answers, 52% have at least one exact match, 80% have one (or more) partial match, and 92% have at least one non-GT match. This shows saturation at top $k = 5$ with a satisfying result of 72% exact matches and 100% factually correct statements, which is a solid performance given the underlying complexity of the QA task.

Episode 3: capability to adopt knowledge bases

The third episode was concerned with the question of whether the system can adapt to different knowledge bases (see DR4). An A/B setup with two identical versions of our QAS but different database versions served as the test framework. The initial version of CORD-19 was from 06/2020, and the more up-to-date one was from 05/2022. During the test, both versions processed similar questions. As an illustration, we asked: “*What type of vaccines are there against COVID-19?*” Based on the 06/2020 version, the most likely answer was: “There are currently no approved vaccines against the COVID-19 virus infection.” – While this is correct in this context, it does not reflect updated knowledge. Answers on ranks two and three argued: “There are no approved specific therapies or vaccines against COVID-19.” Again, the answers were correct but only in the context of a certain time frame. The top answers that draw on the 05/2022 version include: “RNA vaccines” (Park et al., 2021), “COVID-19 vaccines are effective in preventing COVID-19,” and “mRNA vaccines, protein subunit vaccines, and vector vaccines” (see Appendix B). These examples demonstrate how the artifact is capable of incorporating new knowledge, here from scientific sources. However, it has to be noted that the artifact supports such features but requires updated knowledge (databases).

Episode 4: experiment (A/B test)

Besides, we tested the QAS within a naturalistic setting in which over 100 subjects solved a search task. We compared our QAS with the de facto standard systems *Google Search* (general public) and *PubMed* (experts) to observe how the systems perform (see details in Appendix C).

Experimental procedure: subjects, experiment design, and tasks

Our experiment differentiates two *subject groups*, each with treatments concerning the IR systems used for the tasks. Group A represents users without a medical background. It includes subjects from all demographics with only one filter, namely fluent in English. Due to Google’s position in search engines (e.g., market share of over 92%, StatCounter, 2022), we assume that Google is the most frequently used system when searching for information why Group A compares COVID-QAS with Google. Group B represents power users with a biomedical background. It comprises subjects from all demographics that are fluent in English and have completed or are currently enrolled in at least one of the following subjects: Biochemistry, Biological Sciences, Biology, Biomedical Sciences, Dentistry, Genetics, Health and Medicine, Medicine, Nursing, and Pharmacology. Interviews and literature confirmed that PubMed is the most commonly used IR system to retrieve scientific health literature (White, 2020) wherefore Group B compares COVID-QAS against PubMed.

We chose a *within-group experiment* (Field & Hole, 2002) because it is less time-/cost intensive, more sensitive to experimental manipulation, and allows for comparative feedback. All subjects experienced both conditions in our experiment using both systems sequentially. The only difference is that Group A uses Google and Group B PubMed. For randomization, we

employed counterbalancing, where equally sized groups complete each condition (Jhangiani et al., 2019). Two conditions were randomized, namely the system order and the questions for the search tasks. Fifty percent first used COVID-QAS, and the other 50% of users began with Google or PubMed (see Figure 6).

The search task originated from a pool of 20 questions. Each subject answered five questions per system (i.e., ten per run). Each run required two out of four possible question sets. Three question difficulty levels were differentiated; created in a panel to exclude selection bias. Each set comprised two *easy* questions for general knowledge and topics discussed in the mass media (e.g., “What kind of test can diagnose COVID-19?”). *Medium*-difficulty questions require specialized knowledge, but casual users can interpret the answers because a date, numbers, or individual terms are retrieved (e.g., “What is the case fatality rate from SARS and MERS?”). *Difficult* questions involved specific terminology and the answers were elusive without a biomedical background (e.g., “What regulates the secretion of proinflammatory cytokines?”).

Measurement (constructs), data collection, and data processing

To measure the DRs – focus on DR1-DR3 as DR4 was evaluated in episode 3 – we defined items per each of the 11 constructs: **DR1** for reduced cognitive effort is operationalized through information’s quantity, complexity, novelty, and diversity & ambiguity; as an example for quantity: “I can use the system to find answers from the high volume of information about COVID-19;” **DR2** for increased answer quality is studied along the dimensions of relevance, conciseness, justification, and coherence (Breck et al., 2000); **DR3** for minimized knowledge prerequisites is evaluated via ease of use and accessibility of biomedical information (Alexandre et al., 2018). Also, we explored the artifact’s **usefulness** through the overall

user **satisfaction** (Davis, 1989) as well as collected demographic information describing the subjects to ensure a representative sample (see the complete list in Appendix C1).

After validating and refining our experiment design through pretests (see Appendix C2), the data collection started. An online survey was used for the primary study along with the six procedure steps (see Figure 6) (between 06/2022 and 07/2022). Microsoft Forms was used as it fulfills our needs, such as question categories, randomized questions, and an intuitive UI. After finishing the collection, the data was integrated (e.g., downloading and merging files). Then, data cleaning was performed to improve the quality according to five criteria: nonresponses participants with two failed attention checks (Peer et al., 2017), speeding participants (anonymously short response time, Curran, 2016), repetitive responses, and multivariate outliers (e.g., outlier response patterns).

Experiment results: descriptive analysis

Based on an initial set of 122 responses, we applied our data-cleaning procedure leading to 109 participants who completed the survey (see Figure 7).

Experiment results from group A (COVID-QAS vs. Google)

Group A covers 59 non-medical participants from 12 countries (see Appendix C5). Comparing the design instantiated by our artifact – COVID-QAS – to Google (see Figure 8), several main observations have emerged as described below.

Our QAS achieves better mean scores than Google in all constructs of **DR1** ($\Delta = .55$) for reducing users’ cognitive efforts. Nonetheless, participants perceived both systems as positive when searching for pandemic information. The highest discrepancy can be found regarding *complexity*, which indicates that participants

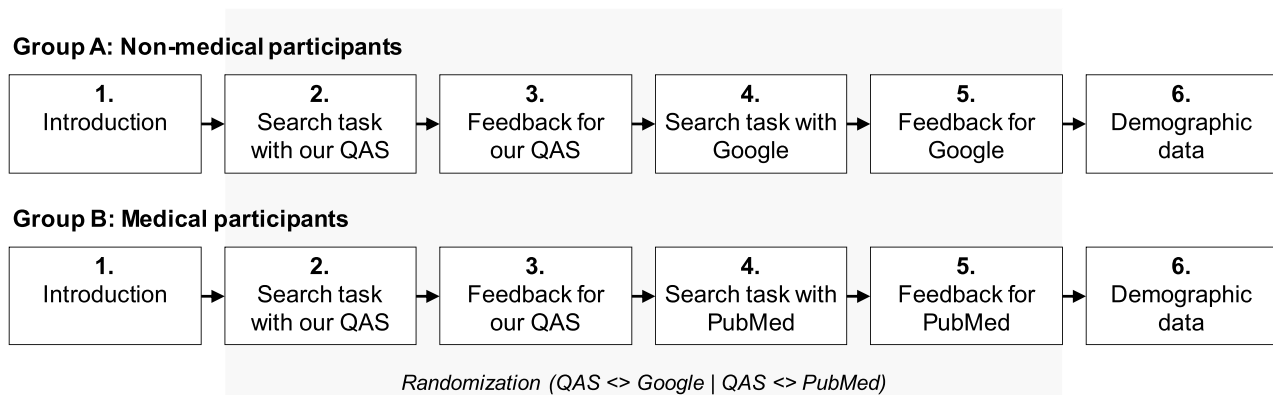


Figure 6. Experimental procedure.

	Group A (non-medical background)	Group B (medical background)	Total
Source: Prolific	31	49	80
Source: SurveyCycle	24	0	24
Source: Personal	4	1	5
	59	50	109

	Group A	Group B	Difference
Age: Median	25	24	1
Age: Mean	26.49	25	1.49
Search engine affinity: Median	6	6	0
Search engine affinity: Mean	5.83	5.62	0.21
IT affinity: Median	6	6	0
IT affinity: Mean	5.86	5.66	.20

	Group A	Group B
Gender:	Female 51 % Male 46%	Female 78% Male 22%
Education (degree):	Bachelor 54% Master 15 % High school 19%	Bachelor 42% Master 22 % High school 28%

Figure 7. Experiment participants and groups.

Group A	Construct mean			Design requirement		
	COVID-QAS	Google	Δ	COVID-QAS	Google	Δ
DR1: Quantity dimension	6.02	5.57	.44	5.72	5.18	.55
DR1: Complexity dimension	5.81	4.83	.97			
DR1: Novelty dimension	5.52	5.49	.03			
DR1: Diversity dimension	5.56	4.81	.75			
DR2: Relevance dimension	5.73	5.18	.56	5.85	4.91	.94
DR2: Conciseness dimension	5.94	4.56	1.37			
DR2: Justification dimension	5.77	4.77	1			
DR2: Coherence dimension	5.96	5.14	.82			
DR3: UI interaction dimension	5.63	5.84	-.21	5.56	5.69	-.14
DR3: Domain dimension	5.48	5.55	-.06			
User satisfaction dimension	5.56	5.20	.35	5.56	5.20	.35

Figure 8. Comparison of the mean for items, constructs, and DRs in group A.

struggle to use Google for complex search tasks in particular. A 27-year-old German female confirmed this: “The more complex the questions, the more difficult they are to Google, while the effort for COVID-QAS remains the same with increasing complexity.” Also, *diversity & ambiguity* highly differs, which points to the fact that Google might be recognized as less suitable for comparing different sources of information.

COVID-QAS exceeds Google’s rating of **DR2** ($\Delta = .94$), so users perceived that the answers have higher quality. While there is a high *relevance*, a 26-year-old German male complained that “the system could not really answer my custom question.” Among the highest differences in favor of COVID-QAS are *conciseness* and *justification*, indicated by feedback including “precise answers from reliable sources.” A reason might be the use of purely scientific databases, while Google searches

the entire web. However, we found potential for improvement, like “the reasoning could be more explained why it is the most similar answer, etc.”

Google outperforms our QAS concerning **DR3** ($\Delta = -.14$) for knowledge prerequisites. Its overall highest-ranked construct for *UI interaction* leads to the largest difference to COVID-QAS ($\Delta = -.21$). The advantages at Google are supported by the qualitative statements, such as COVID-QAS “took ages too [sic!] load.”

User satisfaction was rated on a one to seven Likert scale (see Appendix C2), with our QAS being higher compared to Google ($\Delta = .35$), which is supported by a 23-year-old Mexican male who argued: “The system is easy to use and understand, compared to other search [sic!] engines, it shows you directly the answer to your question and not places where you would have to look for those answers.” This points to the advantage of QA over classical IR. In contrast, a 56-year-old Canadian woman criticizes both systems: “The COVID-QAS results (frankly, like the Google results) are heavily skewed toward ‘official’ government sources.” The comment stressed that even systems with a scientific knowledge base might be insufficient for some users. QAS cannot eliminate the general skepticism toward science

and public authorities. It can only add value if users are open to the technology and have a basic level of trust.

We applied a paired t-test to explore the result’s significance (Ross & Willson, 2017) (see Table 3, aggregated to DRs). The low *p*-values of DR1 and DR3 indicate that COVID-QAS performs better than Google with very high confidence (98% and >99.9%). The difference is highly significant for DR2, so the respondents perceive the answer quality of our QAS as better.

Experiment results from group B (COVID-QAS vs. PubMed)

Group B consists of 50 participants from 15 countries with medical backgrounds, such as Medicine (24%) and Health (16%) (see Appendix C3). Comparing COVID-QAS to PubMed (see Figure 9), the following main observations have emerged: COVID-QAS performs in all **DR1** constructs on reduced cognitive efforts better than PubMed ($\Delta = 1.49$). Particularly the means for *quantity* and *complexity* are higher. The lowest but still better results were achieved concerning *novelty*, where PubMed performs best. *Diversity & ambiguity* point to a major discrepancy; while COVID-QAS has the highest overall mean score, it is the lowest for PubMed.

Table 3. Paired t-test aggregated for the DRs (Group A).

Group A	\bar{X}	SD	SEM	COR	95% CI	t	d	p
DR1	.55	1.68	.23	.05	[0.09, 1.01]	2.4	0.33	.02
DR2	.94	1.66	.23	.04	[0.49, 1.39]	4.16	0.57	<.001
DR3	-.14	1.54	.21	.21	[-0.56, 0.29]	-.66	-.09	.51

mean (\bar{X}), standard deviation (SD), standard error of the mean (SEM), correlation (COR), confidence interval (CI) at 95%, the t-statistic, Cohen’s *d*, two-tailed *p*-value.

Group B	Construct mean			Design requirement		
	COVID-QAS	PubMed	Δ	COVID-QAS	PubMed	Δ
DR1: Quantity dimension	6.04	4.76	1.28	5.98	4.49	1.49
DR1: Complexity dimension	5.93	4.12	1.81			
DR1: Novelty dimension	5.85	5.17	.68			
DR1: Diversity dimension	6.11	3.9	2.21			
DR2: Relevance dimension	5.83	4.28	1.55	5.9	4.44	1.46
DR2: Conciseness dimension	5.95	3.49	2.46			
DR2: Justification dimension	5.73	5.90	-.17			
DR2: Coherence dimension	6.10	4.06	2.01			
DR3: UI interaction dimension	6.32	3.96	2.36	6.18	3.6	2.58
DR3: Domain dimension	6.05	3.24	2.81			
User satisfaction dimension	5.74	4.16	1.58	5.74	4.16	1.58

Figure 9. Comparison of the mean for items, constructs, and DRs in group B.

Concerning **DR2** (quality, $\Delta = 1.46$), the respondents perceived the answers of COVID-QAS as more *relevant* than PubMed. However, a 31-year-old female dentist from the UK mentioned that COVID-QAS is “[...] much simpler and less time consuming than PubMed although answers don’t always quite match the question being asked.” Thus, there is potential for improving the query translation. Also, COVID-QAS scores higher on *conciseness*, which is not surprising because PubMed does not have a QA functionality, and users need to search the entire retrieved documents. Contrarily, we found that PubMed got higher scores for *justification*, which could be attributed to the fact that it benefits from its standing as an established biomedical IR system and its supporters, like the US National Institute of Health.

The greatest difference between the systems can be observed regarding **DR3** (knowledge prerequisites) in favor of COVID-QAS ($\Delta = 2.58$). Especially the *UI interaction* outperforms PubMed. In the words of a 27-year-old female physician from Germany: “COVIDQAS is so much easier and quicker to use than PubMed!”

Also, the overall **user satisfaction** with COVID-QAS is higher ($\Delta = 1.58$). For illustration, a 45-year-old female biological scientist from Nigeria emphasized that “Covid-QAS is a fantastic search engine.” and a 21-year-old South African female noted that “the system is very helpful.” The means further indicate that the artifact is more useful than PubMed in answering questions about COVID-19.

Similar to Group A, we performed a paired t-test (see Table 4). Each *p*-value is less than .001, so COVID-QAS satisfies the three DRs highly significantly better than PubMed. All DRs show a large effect size above the $d > .8$ threshold (Cohen 1988).

Discussion

Ågerfalk et al. (2020) emphasized the central role of IS in alleviating pandemic situations and called for more research on designing useful IS artifacts. By responding to this call, we report on our QAS solution, a system for COVID-19. We build upon the cognitive model of IR interaction and technologies from NLP and neural QAS

to derive knowledge that serves as a blueprint to inform the design of IR systems. Although our artifact has a specific focus, it is intended to provide knowledge for the entire class of QASs supporting different pandemics. Our prototypical implementation yields promising evaluation results in supporting both medical and non-medical users to find answers. This article makes several contributions, which we discuss next.

Implications for misinformation during the pandemic

The general spread of false information has negative effects on the broader society (Moravec et al., 2019; Schoormann et al., 2025; Wei et al., 2022). The exceptional situation faced during the COVID-19 pandemic has accelerated these effects and led to a virus of misinformation (Beisecker et al., 2022; O’Connor & Murphy, 2020; Shirish et al., 2021). Both academic and public communities were confronted with an overwhelming amount of information resulting in a state of “information overload” (Brainard, 2020). Many solutions have emerged over the last years to overcome these problems. These range from human-based to fully automatic detection of false information (Schuetz et al., 2021). We followed the socio-technical nature of IS as managing the information needs presumes the integration of technical, task-related, and domain components. This allows us to take into account the interaction between users with an individual need for information or approach to solving a problem and the knowledge bases provided in (social) media. With our QAS capable of interacting in natural language and providing orientation about the actual context/source of information, we complement existing IS research focusing on, for instance, the representation of pandemic information (e.g., Pietz et al., 2020; Recker, 2021) and visualization of its truthfulness (e.g., Schuetz et al., 2021).

Given that our tool seeks to be accessible and understandable for a larger community, this has implications for ongoing endeavors in the realm of fostering social inclusion (e.g., Schoormann & Kutzner, 2020). The

Table 4. Paired t-test aggregated for the DRs (group B).

Group B	\bar{X}	SD	SEM	COR	95% CI	t	d	p
DR1	1.49	1.76	.25	-.01	[.99, 2.01]	5.94	.85	<0.001
DR2	1.46	1.74	.25	.10	[.97, 1.97]	5.89	.84	<0.001
DR3	2.58	1.68	.24	.12	[2.11, 3.07]	10.78	1.54	<0.001

mean (\bar{X}), standard deviation (SD), standard error of the mean (SEM), correlation (COR), confidence interval (CI) at 95%, the t-statistic, Cohen’s *d*, two-tailed *p*-value.

experiment results point to benefits for both groups of medical experts and laypersons.

Implications for the design of IR systems

The extracted design knowledge helps academia and practice in implementing purposeful IR systems. Comparable to other studies that also draw on Ingwersen's (1996) cognitive model, we can observe both similarities and differences (e.g., Seidel et al., 2008; Sturm & Sunyaev, 2019). For instance, Sturm and Sunyaev (2019) designed a search system for literature and thereby proposed three meta-requirements for *comprehensiveness* (all relevant literature is covered), *precision* (high precision), and *reproducibility* (reliable and transparent search). These requirements and the translated principles share similarities with our research but from a different angle. While multi-sourcing is important in literature reviews to ensure that all relevant items are covered, our QAS uses multiple sources (DP4) to reflect the interdisciplinary nature and the reliability of information (similar data collected from more sources might indicate trueness). As another aspect, flexibility is emphasized to enable the creation of individual literature search strategies. In our case, individual aspects mostly refer to the fact that people have different characteristics that need to be considered. Contrary to other IR systems, we argue that contextual embedding is one of the major advantages. It allows users to understand the actual context and provides a list of most likely answers instead of (indirectly) claiming that a particular answer is a single truth. Our system shows discourses in case there is still uncertainty regarding a question; fair reporting of results.

The situational artifact was available as a free web app, and the code was accessible on GitHub. The relevance for users can be seen from the traffic numbers, as according to Cloudflare's data, about 7,000 different users accessed the site in 06/2022 (i.e., the date when the QAS was online) and 3,500 in 06/2023 without marketing activities.

Implications for the broader QAS research

A particular class of artifacts is addressed, namely QAS. Our study's requirements are in line with prior research on QAS from various domains. For instance, in education authors stressed that among other features systems should be easy to use to find information, apply filters from NLP to disclose relevant information, and be adaptable to the context (Wambsgansß et al., 2021).

It is important to note that extractive QASs, in comparison to generative AI tools (Feuerriegel et al., 2024),

are fundamentally different in their goal setting. Tools based on generative AI, such as ChatGPT, generate text that is aimed at satisfying the user by sounding plausible and human-like. However, the generated text does not need to be factually correct, resulting in so-called hallucinations. Our QAS is a retrieval system, not a generative one, wherefore hallucinations should not occur per design. In addition, generative AI systems with billions of parameters are expensive to train and thus are not updated frequently and, as a result, can fail to capture recent medical advances during a health crisis. In their nature, generative AI systems are not extractive retrieval systems, such as PubMed or Google Search. As a consequence, we excluded these systems as candidates for comparison in our evaluation.

Limitations and future directions

This research is not free of limitations. First, since AI technology evolves rapidly, the technological base for the QAS is subject to constant change; our work was mainly conducted between 2020 and 2022. For example, augmented language models might improve the performance. Second, while we investigated the artifact in multiple evaluation episodes, the next steps should explore its *proof-of-use* (Iivari et al., 2021; Nunamaker et al., 2015). Also, demonstrating how well our QAS can ensure adaptability in a dynamic knowledge environment needs additional work. We conducted experiments with different backup versions of the retrieved knowledge base before and after COVID-19 vaccines were introduced to assess how well the quality of responses improved over time (see Appendix B). Our initial results show how the system reflects the respective state of the art of the underlying database but future research needs to discuss this aspect in more detail. Lastly, our QAS incorporates an extensive collection of research papers from the CORD-19 database, covering various domains. Nonetheless, probably not all information demands, especially critical questions from pandemic skeptics, might be successfully and satisfactorily answered by our QAS. Thus, additional trusted data sources could be integrated, such as those from the WHO. Further data sources can be easily integrated because the innovative learning-based architecture of our neural QAS allows for the expansion of the knowledge base through transfer learning capabilities.

Conclusion

Facing pandemic situations in which false information is spread like a virus itself, this article set out to build an artifact to overcome the increasing overload of

information and derive how-to-knowledge to guide the design of this class of artifacts. Pandemics are hard to handle by the broad society why reliable information needs to be accessible and understandable. A situation that is also faced by (scientific) experts who are concerned with a growing amount of papers. Following a DSR method, we present a neural QAS that employs recent Transformer-based language models to query scientific databases for matching documents and extract the answers as concise sentences. The multi-staged evaluation points to promising technical results (e.g., exact matches, accuracy, F1 score) but also to the usefulness of managing and evaluating information, which is investigated in experiments with over 100 participants. Overall, we hope to complement existing research on detecting false information and designing purposeful artifacts that help users beyond their disciplinary background to handle situations of uncertainty.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Johannes Graf is a senior AI product manager at Drooms AG in the context of real estate and M&A transactions. He received his diploma of Information Systems from the Technische Universität Dresden, Germany. He led the development of AI-based systems for information retrieval, risk analysis, translation, question answering, OCR, and data extraction.

Gino Lancho was at the time of this research endeavor with the Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany.

Kai Heinrich is an Assistant Professor with the Otto-von-Guericke-Universität Magdeburg. His research is at the intersection of artificial intelligence, decision support systems, and human-computer interaction. Research core topics include designing AI-based decision support systems and the study of interactions between humans and AI-based systems. He has authored scholarly publications in international journals in information systems, such as *Business & Information Systems Engineering*, *Decision Support Systems*, and *Electronic Markets*, as well as in various conference proceedings, including ICIS, ECIS, AMCIS, and HICSS.

Frederik Möller is a Junior Professor at TU Braunschweig and a researcher at the Fraunhofer Institute for Software and Systems Engineering, Germany. He holds a Doctorate of Engineering from TU Dortmund University. His research interests include data ecosystems, data space design, inter-organizational data sharing, and methodological foundations of design science research. Frederik's works have been published in several peer-reviewed conferences, such as ICIS or ECIS, and journals, such as *Electronic Markets*, *Business and Information Systems Engineering*, or *IEEE Transactions on*

Engineering Management. In his spare time, he enjoys listening to metalcore with Thorsten.

Thorsten Schoormann is an Associate Professor in the Department of People and Technology, Sustainable Digitalization, at Roskilde University in Denmark and a researcher at the Fraunhofer Institute for Software and Systems Engineering in Germany. Thorsten's interests lie in digital sustainability, digital innovation, and design science research aimed at creating social impact. His work has been published in academic journals including the *Journal of Management Information Systems*, *Information Systems Journal*, *European Journal of Information Systems*, *Decision Support Systems*, and *Business & Information Systems Engineering*. His research primarily got the moves from listening to and reflecting on metal music with Frederik.

Patrick Zschech is a Professor of Business Information Systems, esp. Intelligent Systems and Services at TU Dresden. In addition to his academic career, he worked at Robotron Datenbank-Software GmbH as a developer and an instructor for data science qualification programs. Patrick's research focuses on business analytics, machine learning, and artificial intelligence, with a particular emphasis on designing, analyzing, and applying intelligent information systems. His work has been published in leading OR and IS journals, including *Health Care Management Science*, *European Journal of Operational Research*, *Decision Support Systems*, *Business & Information Systems Engineering*, and *Electronic Markets*, and has been presented at major international conferences such as ICIS, ECIS, and HICSS.

ORCID

Kai Heinrich  <http://orcid.org/0000-0002-4907-6802>

Frederik Möller  <http://orcid.org/0000-0001-6274-701X>

Thorsten Schoormann  <http://orcid.org/0000-0002-3831-1395>

Patrick Zschech  <http://orcid.org/0000-0002-1105-8086>

References

- Abbasiantaeb, Z., & Momtazi, S. (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. *WIREs Data Mining and Knowledge Discovery*, 11(6), e1412. <https://doi.org/10.1002/widm.1412>
- Abdi, A., Idris, N., & Ahmad, Z. (2018). QAPD: An ontology-based question answering system in the physics domain. *Soft Computing*, 22(1), 213–230. <https://doi.org/10.1007/s00500-016-2328-2>
- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16(1), 3–9.
- Ågerfalk, P. J., Conboy, K., & Myers, M. D. (2020). Information systems in the age of pandemics: COVID-19 and beyond. *European Journal of Information Systems*, 29(3), 203–207. <https://doi.org/10.1080/0960085X.2020.1771968>
- Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*:

- Management Information Systems*, 25(1), 107–136. <https://doi.org/10.2307/3250961>
- Alexandre, B., Reynaud, E., Osiurak, F., & Navarro, J. (2018). Acceptance and acceptability criteria: A literature review. *Cognition, Technology & Work*, 20(2), 165–177. <https://doi.org/10.1007/s10111-018-0459-1>
- Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems. *A Survey*, 2(3), 1–12.
- Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., & Gupta, M. (2021). COBERT: COVID-19 question answering system using BERT. *Arabian Journal for Science & Engineering*, 48(8), 11003–11013. <https://doi.org/10.1007/s13369-021-05810-5>
- Bawden, D., & Robinson, L. (2020). *Information overload*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-1360>
- Beisecker, S., Schlereth, C., & Hein, S. (2022). Shades of fake news: How fallacies influence consumers' perception. *European Journal of Information Systems*, 33(1), 41–60. <https://doi.org/10.1080/0960085X.2022.2110000>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat. *Science*, 13(10). <https://doi.org/10.1126/science.abc7839>
- Breck, E., Burger, J. D., Ferro, L., Hirschman, L., House, D., Light, M., & Mani, I. (2000). How to evaluate your question answering system every day and still get real work done. *CoRR*, Cs.Cl/0004008.
- Breja, M., & Jain, S. K. (2022). A survey on non-factoid question answering systems. *International Journal of Computers and Applications*, 44(9), 830–837. <https://doi.org/10.1080/1206212X.2021.1949117>
- Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation*. University of Oxford.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Burton-Jones, A., & Straub, D. W. (2006). Reconceptualizing system usage: An approach and empirical test. *Information Systems Research*, 17(3), 228–246. <https://doi.org/10.1287/isre.1060.0096>
- Casero-Ripolles, A. (2020). Impact of covid-19 on the media system. Communicative and democratic consequences of news consumption during the outbreak. *Profesional de la información*, 29(2). <https://doi.org/10.3145/epi.2020.mar.23>
- Choi, Y., Chiu, C. Y. -I., & Sontag, D. (2016). Learning low-dimensional representations of medical concepts. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, San Francisco, CA, USA (pp. 41–50).
- Cohen, J. (1988). *The effect size. Statistical power analysis for the behavioral sciences* (pp. 77–83). Routledge.
- Cuan-Baltazar, J. Y., Muñoz-Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., & Soto-Vega, E. (2020). Misinformation of COVID-19 on the internet: Infodemiology study. *JMIR Public Health and Surveillance*, 6(2), e18444. <https://doi.org/10.2196/18444>
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- Da Silva, J. W. F., Venceslau, A. D. P., Sales, J. E., Maia, J. G. R., Pinheiro, V. C. M., & Vidal, V. M. P. (2020). A short survey on end-to-end simple question answering systems. *Artificial Intelligence Review*, 53(7), 5429–5453. <https://doi.org/10.1007/s10462-020-09826-5>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Deepset. (2021a). *Deepset/Sentence_bert*. https://huggingface.co/deepset/sentence_bert/tree/main
- Deepset. (2021b). *Deepset-Ai/haystack*. <https://github.com/deepset-ai/haystack>
- Deepset. (2021c). *Haystack documentation - document store*. <https://haystack.deepset.ai/docs/latest/documentstore>
- Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint*.
- Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2021). COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ Digital Medicine*, 4(1), 68. <https://doi.org/10.1038/s41746-021-00437-0>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Field, A., & Hole, G. (2002). *How to design and report experiments*. Sage Publications.
- Gamma, E., Helm, R., Johnson, R. E., & Vlissides, J. (2016). *Design patterns: Elements of reusable object-oriented Software*. Addison-Wesley professional computing series. Pearson Education.
- Ghasemaghahi, M. (2017). The impact of big data on firm data diagnosticity: Mediating role of data quality. *Proceedings of the 38th International Conference on Information Systems*, Seoul, South Korea.
- Ghebreyesus, T. A. (2020). *Virtual press conference on COVID-19 - 11 March 2020*. https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-press-conference-full-and-final-11mar2020.pdf?sfvrsn=cb432bb3_2
- Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly: Management Information Systems*, 19(2), 213–236. <https://doi.org/10.2307/249689>
- Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). The anatomy of a design principle. *Journal of the Association for Information Systems*, 21(6), 1622–1652. <https://doi.org/10.17705/1jais.00649>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly: Management Information Systems*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>

- Hiltz, S. R., & Turoff, M. (1985). Structuring computer-mediated communication systems to avoid information overload. *Communications of the ACM*, 28(7), 680–689. <https://doi.org/10.1145/3894.3895>
- Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: The view from here. *Natural Language Engineering*, 7(4), 275–300. <https://doi.org/10.1017/S1351324901002807>
- Hornik, R., Kikut, A., Jesch, E., Woko, C., Siegel, L., & Kim, K. (2021). Association of COVID-19 misinformation with face mask wearing and social distancing in a nationally representative US sample. *Health Communication*, 36(1), 6–14. <https://doi.org/10.1080/10410236.2020.1847437>
- Huang, Z., Xu, S., Hu, M., Wang, X., Qiu, J., Fu, Y., Zhao, Y., Peng, Y., & Wang, C. (2020). Recent trends in deep learning based open-domain textual question answering systems. *Institute of Electrical and Electronics Engineers Access*, 8, 94341–94356. <https://doi.org/10.1109/ACCESS.2020.2988903>
- Iivari, J., Rotvit Perlt Hansen, M., & Haj-Bolouri, A. (2021). A proposal for minimum reusability evaluation of design principles. *European Journal of Information Systems*, 30(3), 286–303. <https://doi.org/10.1080/0960085X.2020.1793697>
- Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham London.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3–50. <https://doi.org/10.1108/eb026960>
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3), 685–695.
- Jhangiani, R. S., Chiang, I. -C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology*. Kwantlen Polytechnic University.
- John, B., Chua, D., Goh, A., & Wickramasinghe, N. (2016). Graph-based cluster analysis to identify similar questions: A design science approach. *Journal of the Association for Information Systems*, 17(9), 590–613. <https://doi.org/10.17705/1jais.00437>
- John, B., Wickramasinghe, N., & Kurian, J. (2018). Identifying similar questions in healthcare social question answering: A design science research. *Proceedings of the 24th Americas Conference on Information Systems*, New Orleans, USA.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech Recognition*. Prentice hall series in artificial intelligence. Pearson Prentice Hall.
- Kratzwald, B., Eigenmann, A., & Feuerriegel, S. (2019). *RankQA: Neural question answering with answer re-ranking*.
- Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems*, 17(5), 489–504. <https://doi.org/10.1057/ejis.2008.40>
- Laato, S., Najmul Islam, A. K. M., Nazrul Islam, M., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, 29(3), 288–305. <https://doi.org/10.1080/0960085X.2020.1770632>
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A lite BERT for self-supervised learning of language representations*.
- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 25(9), 1248–1258. <https://doi.org/10.1093/jamia/ocy072>
- Li, C. -Y. (2017). Why do online consumers experience information overload? An extension of communication theory. *Journal of Information Science*, 43(6), 835–851. <https://doi.org/10.1177/0165551516670096>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251–266. [https://doi.org/10.1016/0167-9236\(94\)00041-2](https://doi.org/10.1016/0167-9236(94)00041-2)
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6
- Meth, H., Mueller, B., & Maedche, A. (2015). Designing a requirement mining system. *Journal of the Association of Information Systems*, 16(9), 799–837. <https://doi.org/10.17705/1jais.00408>
- Mezzetti, D. (2021a). *NeuML/bert-small-Cord19*. <https://huggingface.co/NeuML/bert-small-cord19>
- Mezzetti, D. (2021b). *NeuML/bert-small-Cord19qa*. <https://huggingface.co/NeuML/bert-small-cord19qa>
- Mohammed, M., Sha'aban, A., Jatau, A. I., Yunusa, I., Isa, A. M., Wada, A. S., Obamiro, K., Zainal, H., & Ibrahim, B. (2022). Assessment of COVID-19 information overload among the general public. *Journal of Racial and Ethnic Health Disparities*, 9(1), 184–192. <https://doi.org/10.1007/s40615-020-00942-0>
- Möller, F., Hansen, M. R., & Schoormann, T. (2022). Synthesizing a solution space for prescriptive design knowledge codification. *Scandinavian Journal of Information Systems*, 34(2), 1. <https://aisel.aisnet.org/sjis/vol34/iss2/1>
- Möller, F., Schoormann, T., Strobel, G., & Hansen, M. R. P. (2022). Unveiling the cloak: Kernel theory use in design science research. *Proceedings of the International Conference on Information Systems (ICIS)*, Copenhagen, Denmark.
- Möller, T., Reina, A., Jayakumar, R., & Pietsch, M. (2020). COVID-QA: A question answering dataset for COVID-19. *COVID-QA: A question answering*.
- Montani, D., Savale, L., Beurnier, A., Colle, R., Noël, N., Pham, T., Monnet, X., & Humbert, M. (2021). Multidisciplinary approach for post-acute COVID-19 syndrome: Time to break down the walls. *The European Respiratory Journal*, 58(1). <https://doi.org/10.1183/13993003.01090-2021>
- Moravec, P., Minas, R., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes No sense at all. *MIS Quarterly: Management Information Systems*, 43(4), 1343–1360. <https://doi.org/10.25300/MISQ/2019/15505>

- Nasery, M., Turel, O., & Yuan, Y. (2023). Combating fake news on social media: A framework, review, and future opportunities. *Communications of the Association for Information Systems*, 53(1), 833–876. <https://doi.org/10.17705/1CAIS.05335>
- NLM. (2024). *An introduction to health literacy by the National Library of Medicine*. Retrieved March 9, 2025, from <https://www.nlm.gov/guides/intro-health-literacy#:~:text=Nearly%209%20out%20of%2010,understanding%2C%20and%20using%20that%20information>
- Norman, C. D., & Skinner, H. A. (2006). eHealth literacy: Essential skills for consumer health in a networked world. *Journal of Medical Internet Research*, 8(2), e9. <https://doi.org/10.2196/jmir.8.2.e9>
- Nunamaker, J. F., Jr., Briggs, R. O., Derrick, D. C., & Schwabe, G. (2015). The last research mile: Achieving both rigor and relevance in information systems research. *Journal of Management Information Systems*, 32(3), 10–47. <https://doi.org/10.1080/07421222.2015.1094961>
- O'Connor, C., & Murphy, M. (2020). Going viral: Doctors must tackle fake news in the COVID-19 pandemic. *BMJ*, 369. <https://doi.org/10.1136/bmj.m1587>
- Olvera-Lobo, M. -D., & Gutiérrez-Artacho, J. (2010). Question-answering systems as efficient sources of terminological information: An evaluation. *Health Information & Libraries Journal*, 27(4), 268–276. <https://doi.org/10.1111/j.1471-1842.2010.00896.x>
- Park, J. W., Lagniton, P. N., Liu, Y., & Xu, R. -H. (2021). mRNA vaccines for COVID-19: What, why and how. *International Journal of Biological Sciences*, 17(6), 1446–1460. <https://doi.org/10.7150/ijbs.59233>
- Peabody, R. L. (1965). BERTRAM M. GROSS. The managing of organizations: The administrative struggle. Vols. I and II. Pp. xx, 971. New York: Free press of Glencoe, 1964. 19.95. *Annals of the American Academy of Political and Social Science*, 360(1), 197–198. <https://doi.org/10.1177/000271626536000140>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Pennington, R., & Tuttle, B. (2007). The effects of information overload on software project risk assessment. *Decision Sciences*, 38(3), 489–526. <https://doi.org/10.1111/j.1540-5915.2007.00167.x>
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770–780. <https://doi.org/10.1177/0956797620939054>
- Pietz, J., McCoy, S., & Wilck, J. H. (2020). Chasing John Snow: Data analytics in the COVID-19 era. *European Journal of Information Systems*, 29(4), 388–404. <https://doi.org/10.1080/0960085X.2020.1793698>
- Rai, A. (2017). Editor's comments: Diversity of design science research. *MIS Quarterly: Management Information Systems*, 41(1), iii–xviii.
- Rajpurkar, P., Jia, R., & Liang, P. (2018). *Know what you don't know: Unanswerable questions for SQuAD*.
- Rajpurkar, P., Jia, R., & Liang, P. (2020). *The Stanford question answering: Dataset-Leaderboard*. <https://rajpurkar.github.io/SQuAD-explorer/>
- Recker, J. (2021). Improving the state-tracking ability of corona dashboards. *European Journal of Information Systems*, 30(5), 476–495. <https://doi.org/10.1080/0960085X.2021.1907235>
- Robles-Flores, J., & Roussinov, D. (2012). Examining question-answering technology from the task technology fit perspective. *Communications of the Association for Information Systems*, 30, 439–454. <https://doi.org/10.17705/1CAIS.03026>
- Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolette, L. D. (2023). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health*, 31(7), 1007–1016. <https://doi.org/10.1007/s10389-021-01658-z>
- Roetzel, P. G. (2019). Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, 12(2), 479–522. <https://doi.org/10.1007/s40685-018-0069-z>
- Ross, A., & Willson, V. L. (2017). Paired samples T-Test. In *Basic and advanced statistical tests: Writing results sections and creating tables and Figures* (pp. 17–19). Sense Publishers. https://doi.org/10.1007/978-94-6351-086-8_4
- Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance. *MIS Quarterly: Management Information Systems*, 43(3), 695–719. <https://doi.org/10.25300/MISQ/2019/13747>
- Saunders, C., Wiener, M., Klett, S., & Sprenger, S. (2017). The impact of mental representations on ICT-Related overload in the use of mobile phones. *Journal of Management Information Systems*, 34(3), 803–825. <https://doi.org/10.1080/07421222.2017.1373010>
- Schick, A. G., Gordon, L. A., & Haka, S. (1990). Information overload: A temporal approach. *Accounting, Organizations & Society*, 15(3), 199–220. [https://doi.org/10.1016/0361-3682\(90\)90005-F](https://doi.org/10.1016/0361-3682(90)90005-F)
- Schneider, S. C. (1987). Information overload: Causes and consequences. *Human Systems Management*, 7(2), 143–153. <https://doi.org/10.3233/HSM-1987-7207>
- Schoormann, T., & Kutzner, K. (2020). Towards understanding social sustainability: An information systems research-perspective. *Proceedings of the 41st International Conference on Information Systems*, India.
- Schoormann, T., Möller, F., Chandra Kruse, L., & Otto, B. (2024). BAUSTEIN-A design tool for configuring and representing design research. *Information Systems Journal*, 34(6), 1871–1901. <https://doi.org/10.1111/isj.12516>
- Schoormann, T., Möller, F., Hoppe, C., & Vom Brocke, J. (2025). Digital sustainability - understanding and managing tensions. *Business & Information Systems Engineering*. <https://doi.org/10.1007/s12599-025-00937-3>
- Schuetz, S. W., Sykes, T. A., & Venkatesh, V. (2021). Combating COVID-19 fake news on social media through fact checking: Antecedents and consequences. *European*

- Journal of Information Systems*, 30(4), 376–388. <https://doi.org/10.1080/0960085X.2021.1895682>
- Seidel, S., Chandra Kruse, L., Székely, N., Gau, M., Stieger, D., Peffers, K., Tuunanen, T., Niehaves, B., & Lyytinen, K. (2017). Design principles for sensemaking support systems in environmental sustainability transformations. *European Journal of Information Systems*, 27(2), 221–247. <https://doi.org/10.1057/s41303-017-0039-0>
- Seidel, S., Müller-Wienbergen, F., Rosemann, M., & Becker, J. (2008). A conceptual framework for information retrieval to support creativity in business processes. *Proceedings of the 16th European Conference on Information Systems*, Galway, Ireland.
- Shahid, R., Shoker, M., Chu, L. M., Frehlick, R., Ward, H., & Pahwa, P. (2022). Impact of low health literacy on patients' health outcomes: A multicenter cohort study. *BMC Health Services Research*, 22(1), 1148. <https://doi.org/10.1186/s12913-022-08527-9>
- Sharma, Y., & Gupta, S. (2018). Deep learning approaches for question answering system. *Procedia Computer Science*, 132, 785–794. <https://doi.org/10.1016/j.procs.2018.05.090>
- Shirish, A., Srivastava, S. C., & Chandra, S. (2021). Impact of mobile connectivity and freedom on fake news propensity during the COVID-19 pandemic: A cross-country empirical examination. *European Journal of Information Systems*, 30(3), 322–341. <https://doi.org/10.1080/0960085X.2021.1886614>
- Siering, M., Muntermann, J., & Grčar, M. (2021). Design principles for robust fraud detection: The case of stock market manipulations. *Journal of the Association for Information Systems*, 22(1), 156–178. <https://doi.org/10.17705/1jais.00657>
- Sørensen, K., Pelikan, J. M., Röthlin, F., Ganahl, K., Slonska, Z., Doyle, G., Fullam, J., Kondilis, B., Agrafiotis, D., Ueters, E., Falcon, M., Mensing, M., Tchamov, K., van Broucke, S. V. D., Brand, H., & on behalf of the HLS-EU Consortium. (2015). Health literacy in Europe: Comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health*, 25(6), 1053–1058. <https://doi.org/10.1093/eurpub/ckv043>
- StatCounter. (2022). *Search engine market share worldwide*. <https://gs.statcounter.com/search-engine-market-share>
- Sturm, B., & Sunyaev, A. (2019). Design principles for systematic search systems: A holistic synthesis of a rigorous multi-cycle design science research journey. *Business & Information Systems Engineering*, 61(1), 91–111. <https://doi.org/10.1007/s12599-018-0569-6>
- Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., & Fung, P. (2020). CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management.
- Szmuda, T., Özdemir, C., Ali, S., Singh, A., Syed, M. T., & Słoniewski, P. (2020). Readability of online patient education material for the novel coronavirus disease (COVID-19): A cross-sectional health literacy study. *Public Health*, 185, 21–25. <https://doi.org/10.1016/j.puhe.2020.05.041>
- Tawfik, A. A., Kochendorfer, K. M., Saparova, D., Al Ghenaimi, S., & Moore, J. L. (2014). “I don’t have time to dig back through this”: The role of semantic search in supporting physician information seeking in an electronic health record. *Performance Improvement Quarterly*, 26(4), 75–91. <https://doi.org/10.1002/piq.21158>
- Turc, I., Chang, M. -W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*. abs/1908.08962
- Ulmer, A. (2020). *Indian guru’s tips to Ward off coronavirus anger health professionals*. <https://www.reuters.com/article/us-health-coronavirus-india-ayurveda-idUSKBN21515M>
- van der Meer, T. G. L. A., & Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, 35(5), 560–575. <https://doi.org/10.1080/10410236.2019.1573295>
- Vanopstal, K., Buysschaert, J., Laureys, G., & Stichele, R. V. (2013). Lost in PubMed. Factors influencing the success of medical information retrieval. *Expert Systems with Applications*, 40(10), 4106–4114. <https://doi.org/10.1016/j.eswa.2013.01.036>
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), 77–89. <https://doi.org/10.1057/ejis.2014.36>
- Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2), 219–240. <https://doi.org/10.1111/j.1540-5915.1991.tb00344.x>
- Wambsganß, T., Haas, L., & Söllner, M. (2021). Towards the design of a student-centered question-answering system in educational settings. *Proceedings of the 29th European Conference on Information Systems*, Marrakesh, Morocco.
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B. & Kohlmeier, S. (2020). *CORD-19: The COVID-19 open research dataset*.
- Wang, S., Pang, M. -S., & Pavlou, P. A. (2022). Seeing is believing? How including a video in fake news influences users’ reporting of fake news to social media platforms. *MIS Quarterly*, 45(3), 1323–1354. <https://doi.org/10.25300/MISQ/2022/16296>
- Wei, X., Zhang, Z., Zhang, M., Chen, W., & Zeng, D. D. (2022). Combining crowd and machine intelligence to detect false news on social media. *MIS Quarterly*, 46(2), 977–1008. <https://doi.org/10.25300/MISQ/2022/16256>
- White, J. (2020). PubMed 2.0. *Medical Reference Services Quarterly*, 39(4), 382–387. <https://doi.org/10.1080/02763869.2020.1826228>
- WHO. (2020). *Novel Coronavirus(2019-nCoV) situation report-13*. https://www.who.int/docs/default-source/coronavirus/situation-reports/20200202-sitrep-13-ncov-v3.pdf?sfvrsn=195f4010_6
- Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, 55(3), 249–270. <https://doi.org/10.1108/EUM0000000007145>
- Yang, M. -C., Lee, D. -G., Park, S. -Y., & Rim, H. -C. (2015). Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications*, 42(23), 9086–9104. <https://doi.org/10.1016/j.eswa.2015.07.009>

- Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3), 55–75. <https://doi.org/10.1109/MCI.2018.2840738>
- Zhu, D., Nimmagadda, S. L., Wong, K. W., & Reiners, T. (2023). Relevance judgment convergence degree--A measure of assessors inconsistency for information retrieval datasets. In G. C. Silaghi, R. A. Buchmann, V. Niculescu, G. Czibula, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), *Advances in information systems development: AI for is development and operations* (pp. 149–168). Springer International Publishing. https://doi.org/10.1007/978-3-031-32418-5_9
- Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T. -S. (2021). *Retrieving and reading: A comprehensive survey on open-domain question answering*.
- Zschech, P., Horn, R., Hörschele, D., Janiesch, C., & Heinrich, K. (2020). Intelligent user assistance for automated data mining method selection. *Business & Information Systems Engineering*, 62(3), 227–247. <https://doi.org/10.1007/s12599-020-00642-3>