SCM
Studies in Communication and Media

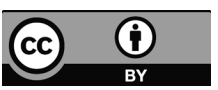
FULL PAPER

Beyond fact-checking: Harnessing AI for news quality assessment

Beyond Fact-Checking: KI-gestützte Qualitätsbewertung im Journalismus

Tobias Schmidt, Jonas Rieger, Jörg Rahnenführer, Astrid Viciano & Holger Wormer

Tobias Schmidt (M. A.), TU Dortmund University, Institute and School of Journalism, Emil-Figge-Straße 50, 44227 Dortmund, Germany. Contact: tobias3.schmidt@tu-dortmund.de. ORCID: https://orcid.org/0000-0001-7330-4943
Jonas Rieger (Dr.), TU Dortmund University, Department of Statistics, Vogelpothsweg 87, 44227 Dortmund, Germany. Contact: rieger@statistik.tu-dortmund.de. ORCID: https://orcid.org/0000-0002-0007-4478
Jörg Rahnenführer (Prof. Dr.), TU Dortmund University, Department of Statistics, Vogelpothsweg 87, 44227 Dortmund, Germany. Contact: rahnenfuehrer@statistik.tu-dortmund.de. ORCID: https://orcid.org/0000-0002-8947-440X
Astrid Viciano (Dr.), TU Dortmund University, Institute and School of Journalism, Emil-Figge-Straße 50, 44227 Dortmund, Germany. Contact: astrid.viciano@tu-dortmund.de.
Holger Wormer (Prof. Dipl.-Chem.), TU Dortmund University, Institute and School of Journalism, Emil-Figge-Straße 50, 44227 Dortmund, Germany. Contact: holger.wormer@tu-dortmund.de. ORCID: https://orcid.org/0009-0005-7708-2934



FULL PAPER

Beyond fact-checking: Harnessing AI for news quality assessment

Beyond Fact-Checking: KI-gestützte Qualitätsbewertung im Journalismus

Tobias Schmidt, Jonas Rieger, Jörg Rahnenführer, Astrid Viciano & Holger Wormer

Abstract: Reliable scientific information is crucial for shaping public opinion and informing economic, social, and health-related decisions. While significant research has focused on detecting misinformation, the automated assessment of the overall quality of journalistic content remains largely unexplored. Addressing this gap, our study investigates whether different artificial intelligence approaches ranging from classical statistical methods ("bag-of-words") to advanced large language models (LLMs) can assist in evaluating the quality of medical news articles based on ten established journalistic quality criteria. Drawing on a dataset of 240 articles and corresponding expert reviews, we show that machine learning models can indeed assess article quality with considerable accuracy. While several models fall short of human-level performance, fine-tuned LLMs perform well and, in most cases, match the level of agreement with a consolidated expert benchmark observed among individual human reviewers. The best-performing model, a fine-tuned GPT-4o model, detects poor-quality articles reliably when evaluated against expert assessments (macro-averaged F1: 0.69 vs. 0.57), suggesting its potential for use in editorial workflows, such as filtering incoming material or double-checking an already written text. Although our study focuses on medical journalism's well-defined criteria, the approach likely holds promise for enhancing quality assurance in other areas of journalism and science communication where similar quality criteria exist.

Keywords: Journalistic quality, science communication, machine learning, large language models, automation

Zusammenfassung: Zuverlässige wissenschaftliche Informationen sind entscheidend für die öffentliche Meinungsbildung sowie für wirtschaftliche, soziale und gesundheitspolitische Entscheidungen. Während sich zahlreiche Studien mit der Erkennung von Falschinformationen befassen, ist die automatisierte Bewertung der Gesamtqualität journalistischer Inhalte bislang weitgehend unerforscht. Diese Studie schließt diese Forschungslücke, indem sie untersucht, inwiefern verschiedene KI-Modelle – von klassischen statistischen Methoden („Bag-of-Words") bis hin zu fortgeschrittenen Large Language Models (LLMs) – die Qualität medizinischer Nachrichtenartikel anhand von zehn etablierten journalistischen Qualitätskriterien bewerten können. Anhand eines Datensatzes mit 240 medizinjournalistischen Artikeln und zugehörigen Expertengutachten zu diesen Texten zeigen wir, dass Machine-Learning-Modelle durchaus in der Lage sind, die journalistische Qualität eines Artikels valide zu beurteilen. Zwar erreichen mehrere Modelle nicht das Leistungsniveau menschlicher Expert*innen, speziell feingetunte LLMs jedoch erzielen in den meisten Fällen ein

Maß an Übereinstimmung mit einer konsolidierten Expertenbewertung, das mit dem Übereinstimmungsgrad zwischen menschlichen Expert*innen vergleichbar ist. Das leistungsstärkste Modell (ein feingetuntes GPT-4o) identifiziert qualitativ schlechte Artikel ebenso verlässlich wie menschliche Gutachter*innen (macro-averaged F1: 0.69 vs. 0.57) und deutet damit auf ein Potenzial für den Einsatz in redaktionellen Workflows hin, etwa zur Filterung eingehender Inhalte oder zur Nachkontrolle bereits verfasster Texte. Obwohl sich die Studie auf Qualitätskriterien aus dem Medizinjournalismus stützt, dürfte sich der Ansatz auch in anderen Bereichen des Journalismus und der Wissenschaftskommunikation anwenden lassen, sofern dort vergleichbare Qualitätskriterien existieren.

Schlagwörter: journalistische Qualität, Wissenschaftskommunikation, maschinelles Lernen, Large Language Models, Automatisierung

1. Introduction

Misinformation has become "something of a catch-all term" (Cacciatore, 2021, p. 1) and has attracted the attention of a broad range of disciplines. As early as 2020, a systematic review focused on news creation and consumption identified 2277 fake news-related scientific studies in the literature (Kim et al., 2021). The need for such studies is more important than ever, as we experience that "fake news is spreading day by day" (Jindal et al., 2023, p. 375). At the same time, several authors have pointed out that fostering trust in media may require more than simply combating misinformation in already published articles (e.g., Acerbi et al., 2022). Sorting out the worst misinformation and the most brazen lies does not automatically mean that the audience of (digital) media gets trustworthy and useful information, as even "seemingly trustworthy news sites push misleading headlines" (West & Bergstrom, 2021, p. 1). Following the idea that interventions should primarily seek to build trust in factual information (Tay et al., 2024), it becomes essential to ensure that information sources themselves consistently meet high-quality standards. In view of the cutbacks in many editorial offices, however, this is becoming an increasing challenge for the remaining staff.

Given these constraints, a growing number of studies have explored the potential of artificial intelligence (AI) tools to support journalistic workflows, especially in detecting and counteracting misinformation (Islam et al., 2020; Kumar et al., 2020). However, despite a growing recognition that a "fighting for information" could be far more effective than merely "fighting misinformation" (Acerbi et al., 2022), the potential of AI to assist with the systematic assessment of overall journalistic quality – beyond mere fact-checking – remains largely underexplored (Liu et al., 2025).

In this article, we therefore want to approach the semi-automated evaluation of news content from the other side of the scale between intentional disinformation (as probably the worst kind of information) and good (that is, trustworthy and useful) information. Using reporting on medical research as a case study, we assess the capability of several machine learning (ML) models, including large language models (LLMs), to evaluate the quality of science news content based on established criteria for good medical journalism. The example of medical research as a scientific topic seems particularly suitable for two reasons: First, re-

porting on medical sciences not only accounts constantly for the largest share of science reporting (Bucchi & Mazzolini, 2003; Clark & Illman, 2006; Summ & Volpers, 2016) but misleading information in this field has also the largest potential of causing harm among media recipients (Schwitzer, 2008). Secondly, the quality criteria for good reporting (partly based on standards known from evidence-based medicine) are probably the most developed and precise in science journalism, if not in journalism at all. The proposed concept as well as the developed models presented in this study could, however, be extended to other kinds of reporting, too, given that an elaborated set of quality criteria and standards has been developed.

In the following, we first provide an overview of established concepts and quality criteria for "good reporting" in general, and for science communication in particular, with a specific focus on medical journalism. Next, we examine recent research on ML applications in journalism and explore the potential of AI tools to contribute to quality assessment beyond fact-checking alone. We then propose a methodological framework for developing and evaluating AI tools specifically designed for quality assessment in journalism. Specifically, we evaluate the performance of two common statistical classification models, several LLMs used in a zero-shot setting – that is, without any task-specific fine-tuning – and three fine-tuned LLMs in autonomously assessing the quality of science news articles based on predefined criteria. By comparing these AI-driven assessments with human expert evaluations, we identify the strengths and limitations of each model. Finally, we discuss the adaptability of our approach and suggest future directions for integrating AI-driven quality assessment into the evolving digital media landscape.

2. Related work

Quality of public information is a broad issue which has been discussed in theory and practice ever since public media existed. Numerous terms have emerged to conceptualise this field, with Denis McQuail's (1993) "media performance" being among the most frequently cited concepts. Other commonly used terms include "journalistic" or "newspaper excellence" (Gladney, 1996), while the phrase "quality in journalism" is often preferred in German-language journalism research. Some frequently mentioned general criteria for quality in the media include *facticity* and *accuracy*, *relevance*, *timeliness*, *independence*, *diversity of topics and sources*, and *usefulness*, among others (Meier, 2019). Many of these criteria are applicable beyond traditional journalism, extending to other communication formats like public relations or social media posts (Meier, 2019) and are also mentioned by science communication researchers (Fähnrich et al., 2023; Silva Luna et al., 2025). However, catalogues and recommendations for communicating specific scientific disciplines to the public often go beyond the general quality criteria mentioned above.

One of the first criteria-based approaches to measure the content quality of reports from a specific scientific discipline can be found in Canada (Oxman et al., 1993). With their index of scientific quality (ISQ) for health reports in the lay

press, Oxman et al. (1993) built on earlier work in the field of the so-called accuracy research, concentrating rather on lists of error categories than on quality criteria (Tankard & Ryan, 1974). A systematic monitoring of the quality of medical news reporting using defined criteria combined with a peer-review system came from the *Media Doctor* project in Australia (Smith et al., 2005; Wilson et al., 2009), followed by *HealthNewsReview.org* in the US with the same ten criteria (Schwitzer, 2008).

Building on this international tradition, the German quality assessment initiative "Medien-Doktor" was introduced in 2010, following the same conceptual approach (Wormer, 2011). It has since been adapted to other communication formats (e.g., press releases) and further developed for additional scientific domains, such as environmental reporting (Rögener & Wormer, 2017). Today, it constitutes the largest publicly available repository of expert-reviewed science news articles in the German context, offering a unique resource for empirical research on science journalism quality. Given its systematic methodology, transparent evaluation scheme, and rich corpus of annotated news articles, the Medien-Doktor framework is particularly well-suited for the present study's aim to benchmark automated models against human expert evaluations. We therefore adopt its ten medicine-related quality criteria as the foundation for our analysis, as summarized in Table 1.

Table 1. Overview of the ten Medien-Doktor quality criteria, adapted and reordered from Schwitzer (2008) and the former HealthNewsReview.org in the US.

Criterion 1	Article adequately explains benefits of a new drug/treatment/procedure...
Criterion 2	Adequately explains potential harms
Criterion 3	Reviews the study methodology or the quality of the evidence
Criterion 4	Seeks out independent sources and discloses potential conflicts of interest
Criterion 5	Appears not to rely solely on a news release
Criterion 6	Establishes the true novelty of the idea
Criterion 7	Compares the new idea with existing alternatives
Criterion 8	Establishes the availability of the product or procedure
Criterion 9	Adequately discusses costs
Criterion 10	Avoids disease mongering

These ten criteria aim to make quality in medical journalism assessable in a structured, transparent manner. Ideally, a high-quality article would fulfil all ten criteria, by clearly stating the benefits (criterion 1) and risks (2) of a new procedure, explaining the evidence base (3), providing information on availability (8), communicating all of this in a sober and factual tone (10), and so forth. In the time-pressured reality of journalistic practice, however, this ideal is rarely achieved. The Medien-Doktor project, much like its international counterparts, therefore, pursues a dual goal: To sensitise journalists to these criteria and to make transpa-

rent for the public which articles merit closer scrutiny. For a detailed rationale and the development process of the criteria, see Wormer (2011) and Schwitzer (2008).

2.1 AI systems in the newsroom

Established quality criteria like the ones presented above hold the potential to assist journalists and science communicators in their everyday tasks. They facilitate the screening of incoming press releases or agency news, help identify weaknesses and low-quality documents and support the detection of the most promising material, which can then serve as a starting point for thorough investigations. At the same time, it is widely acknowledged that most newsrooms face significant time and labour constraints. Screening articles and potentially double-checking already written texts takes time, a resource often missing in modern newsrooms. As a result, time pressure not only leads to less diversity and overall, less cross-checking of articles (Reich & Godler, 2014), but also impacts communicators' ability to effectively fact-check information (Himma-Kadakas & Ojamets, 2022). Some argue that, under such circumstances, journalism is prone to "become churnalism, the recycling of pre-packaged public relations and press agency copy" (Davies, 2008, p. 59) – a challenge that persists even in the presence of well-defined quality criteria.

Given these constraints and external factors such as competitive dynamics and pressure to innovate, media outlets have increasingly turned to automation tools designed to support journalists in their daily work (Hermida & Simon, 2025; Simon, 2024). Commonly referred to as part of *Automated journalism* (Carlson, 2015; Caswell & Dörr, 2018; Diakopoulos, 2019), *Algorithmic journalism* (Dörr, 2016; Kotenidis & Veglis, 2021), or *Computational journalism* (Cohen et al., 2011; Thurman, 2019), these tools typically involve the use of algorithms and data analysis to support journalistic tasks. This can include identifying newsworthy trends or stories within data (Liu et al., 2016; Stray, 2019), producing routine articles on topics like sports or finance, or structuring news websites based on recommender systems (Bodó et al., 2019; Latar, 2018). Other tools assist with analyzing unstructured documents, data wrangling or lead generation. Especially the potential of ML models to detect and counteract misinformation has aroused attention in this context (Islam et al., 2020; Kumar et al., 2020).

With the advent of generative artificial intelligence (GenAI) and particularly LLMs, the implementation of AI tools in science communication in general (Schäfer, 2023) and especially in journalism has gained new momentum (Diakopoulos et al., 2024; Simon, 2024). In this context, the term *journalistic Artificial Intelligence* or *journalistic AI* has gained prominence, often used as an umbrella term for all kinds of technologies, "from machine learning to automated decision making, that are used along the entire news production chain" (Helberger et al., 2022, p. 1606). Modern systems are used to summarize or rephrase texts, to generate headlines, personalize news, or assist in real-time fact-checking and background research. Compared to earlier tools, their accessibility and versatility have made them far more present in everyday journalistic practice. Correspondingly,

expectations have grown that such tools could relieve journalists from repetitive and time-consuming tasks, allowing them to dedicate more time to investigative reporting and complex research (Van Dalen, 2012; Wolf, 2024).

However, while there is considerable enthusiasm surrounding the use of modern journalistic AI in newsrooms, the task of evaluating the overall quality of content is rarely mentioned in this context. When discussing the potential of AI systems to filter articles, e.g., researchers typically elaborate on verification, or claim-matching, rather than systems to check the overall article quality (Simon, 2024). As discussed earlier, this approach might be insufficient for effectively promoting trustworthy information. Perhaps most strikingly, the Journalism AI project from the London School of Economics and Political Science lists numerous case studies from around the world on how AI systems are currently utilised in journalism – and reveals that none of the > 250 presented case studies (as of March 2026, subject to future updates) employ AI tools for the quality assessment of news content (LSE, 2025). Most of the related research and practical applications predominantly focus on news production and misinformation detection. That is, the extent to which AI systems can evaluate or "double-check" the overall quality of media content beyond fact-checking remains unclear.

A rare example of studies discussing automation tools for evaluating (science) news quality rather than detecting fake news was presented by Løvlie et al. (2023). The authors developed and evaluated a "Scientific Evidence Indicator" (SEI) designed to help readers assess the strength of evidence and uncertainty in science journalism articles. Their tool evaluates scientific studies based on measurable criteria such as peer review status and methodological quality. Another related study was put forward by Choi et al. (2021), who trained an artificial neural network (ANN) from scratch to predict audience-related news quality assessment. Combining traditional survey methods with ANN-training and other computational approaches, the authors identified journalistic values such as believability, depth, and diversity as stronger predictors of audience-rated news quality than linguistic features or audience attributes. The study that comes closest to our approach was performed by Liu et al. (2025). The authors used the expert-annotated *HealthNewsReview* dataset to evaluate the capability of traditional ML models and *GPT-3.5-Turbo* to rate the quality of health news across established journalistic criteria. Their findings show that while *GPT-3.5* can generate clear and contextually relevant explanations, its actual rating accuracy remains limited and falls short of the traditional ML models. Thus, the authors demonstrate both the potential and the shortcomings of one specific LLM when applied to multidimensional quality evaluation tasks, leaving room for further research in this direction. Beyond these few studies, however, the potential of ML models to support criteria-based quality evaluation of journalistic content remains largely unexplored.

2.2 LLMs in content classification: Promise, pitfalls, and alternatives

Large language models, in particular modern state-of-the-art (SOTA) models like OpenAI's *GPT-5* or Anthropic's *Claude 4.5 Sonnet* have demonstrated strong performance across a variety of natural language processing (NLP) tasks, including natural language understanding, text classification, and information extraction (Chang et al., 2023; Chiang et al., 2024). Especially their ability to understand and process highly diverse texts and their ease of access position them as an effective tool for evaluating diverse types of articles when no other assistant systems exist. Given these strengths, it may seem intuitive to regard LLMs as a default solution for tasks like the one addressed in this study.

At the same time, the use of LLMs is not without concern. Several researchers highlight risks associated with uncritical adoption of AI (and especially LLM) outputs (Mahony & Chen, 2025; Wu, 2024). The "black-box" nature of LLMs makes it difficult to fully understand the decision-making process behind their outputs, which can lead to ethical and practical challenges in journalism and even provoke distrust and hostility (Bail, 2024; Noain Sánchez, 2022; Schäfer et al., 2024). Furthermore, generative AI tools may display various biases, not only producing offensive content, such as racist or misogynistic language, but also hallucinating inaccurate information that could further exacerbate the already existing challenges of disseminating trustworthy information (Schäfer, 2023). Many researchers even recognize threats to editorial independence and the potential erosion of core journalistic values such as accuracy, transparency, and accountability (Cordero et al., 2025; Møller et al., 2025; Wolf, 2024). Last but not least, the increasing reliance on proprietary LLMs – often developed by a small number of global tech companies – can reinforce structural dependencies and limit the autonomy of newsrooms (Simon, 2024).

Considering these challenges, exploring more transparent and interpretable modeling approaches becomes increasingly important. Targeted ML solutions, such as those designed for specific classification tasks, may offer more reliability and control than general-purpose, free-text-generating systems, especially in classification tasks. Classical statistical methods such as Random Forests (Breiman, 2001) or Naive Bayes classifiers (McCallum & Nigam, 1998) offer greater potential for interpretability, which can be crucial in contexts where transparency and accountability are required. Furthermore, several studies have shown that LLMs do not consistently outperform simpler models across all tasks (Liu et al., 2025; Qin et al., 2023; Ziems et al., 2024). Thus, we see considerable value in not solely evaluating *ChatGPT*'s capabilities at a certain task (as is often the case in communication science research) but also benchmarking it against other model types.

2.3 Objective and research questions

Following these concerns, our research setup includes both classical statistical models and several state-of-the-art LLMs. This complementary approach makes it possible to investigate not only the capabilities of general-purpose language models but also the potential of simpler, more transparent classification systems

for content evaluation (Liu et al., 2025). Additionally, we incorporate fine-tuning of LLMs, utilising both small open-source and large proprietary models. This setup allows us to identify the strengths and limitations of each model – and to test whether the limitations reported by Liu et al. (2025) persist regardless of the choice of model and once models are adapted to domain-specific data.

We ask the following research questions: Using the example of medicine-related news articles, can machine learning models reliably assess article quality? Precisely, how far can these models approach human expert performance in evaluating the content quality based on established quality criteria? (RQ1) What are the minimum technical requirements (in terms of model size and complexity) needed to achieve expert-level accuracy? (RQ2) How does prediction accuracy vary across different quality criteria? (RQ3) To what extent can machine learning models detect articles of low quality, serving as inbound or outbound control? (RQ4) Following these questions, we aim to provide insights into how, and to what extent, AI can assist journalists in assessing the overall quality of various texts, be it an incoming press release, material from a news agency, or their own articles, as a final check before publication.

3. Data and methods

We evaluate the effectiveness of a Random Forest (RF) classifier, a Naive Bayes (NB) model, the zero-shot ("out-of-the-box") capabilities of several LLMs, and the performance of three different LLMs that were fine-tuned (ft) specifically on the present task.

Random Forest and Naive Bayes are two well-established classification algorithms commonly used in text analysis. Naive Bayes is a probabilistic model based on Bayes' theorem that assumes independence between features/words, an assumption that is often violated in natural language but still delivers robust results for classification tasks (Zhang, 2004). Random Forest, in contrast, is an ensemble learning method that combines the predictions of multiple decision trees to improve classification accuracy and control overfitting (Breiman, 2001). Both methods rely on simple word-frequency-based representations of text (commonly referred to as "bag-of-words" assumption), making them particularly suitable for settings that prioritise interpretability, resource efficiency, and methodological transparency.[1] In addition, their simplicity makes inference computationally much less intensive, making them suitable for rapid analysis or applications where computational resources are limited.

As mentioned earlier, LLMs provide a fundamentally different approach to text evaluation, as they operate on semantic representations rather than surface-level word frequencies. While the NB and RF models in our implementation operate on lexical features, LLMs can leverage contextual understanding and reaso-

1 It is important to note that "transparency" in this context typically denotes the potential to examine how model features contribute to predictions, for instance by tracing influential words or analysing interactions between them. Still, performing such analyses requires additional effort and does not necessarily translate into valuable practical insights.

ning to evaluate complex criteria in unstructured text. Furthermore, LLMs are particularly advantageous when annotated data are scarce – which are needed for training a NB or RF model –, as they are pretrained and can simply be instructed in natural language. However, given the findings of Liu et al. (2025), we expect moderate performance levels from these models and hypothesise that the specialized ML models exceed the zero-shot capabilities of certain LLMs, especially older and smaller ones. To evaluate the capabilities of LLMs, we test a variety of models that are chosen for their strong performance on various tasks, as indicated by the Chatbot Arena LLM Leaderboard (Chiang et al., 2024). Our selection includes both small and large models, as well as proprietary and open-source models.

Additionally, we perform LLM fine-tuning, since several researchers demonstrate LLM's potential in coding/annotation tasks when specifically trained (Gilardi et al., 2023; Mellon et al., 2024; Rieger et al., 2024). Fine-tuning makes it possible to adapt models to a specific domain (medical-related media articles in our case) and tailor their responses, which significantly improves model performance, especially in areas where precision and task-specific knowledge are crucial (Wu et al., 2023). Precisely, we test the open-source model *Mistral-7B-Instruct-v0.3* from Mistral AI, which we access via the open-source model library Hugging Face, and two models offered by OpenAI, *GPT-4o* and *GPT-4o-mini*.

An overview of all tested models is presented in Table 2. Supplementary material for this article is available on the Open Science Framework (OSF): https://osf.io/6yk82/.

Table 2. Overview of all tested models.

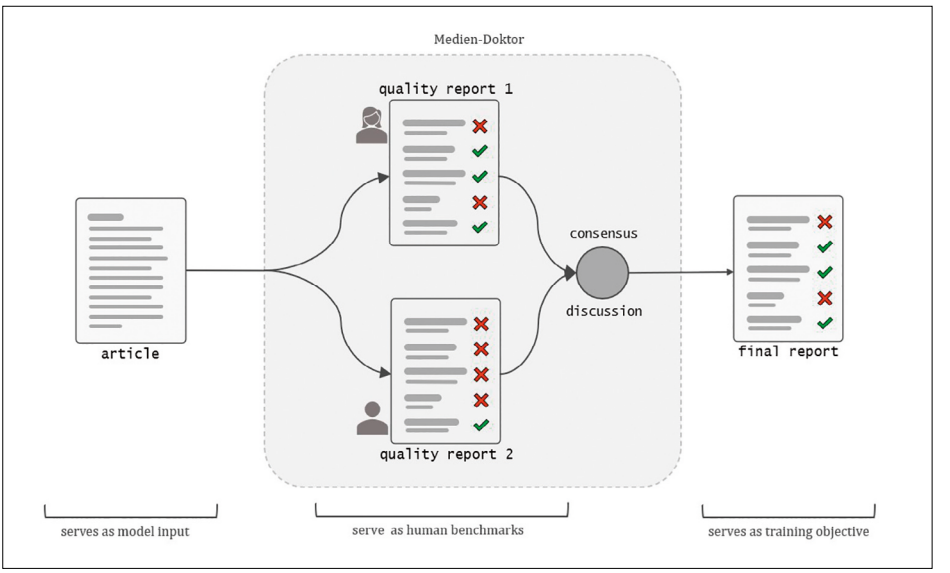
Bag-of-words models	Large language models	
	Zero-shot LLM	Fine-tuned LLM
Naive Bayes Random Forest	GPT-3.5-turbo, GPT-4o, GPT-4-1106-Preview, Claude-3.5-Sonnet Mistral 7B, Mistral 8x22B, Mistral Large Llama 3 8B, Llama 3 70B	GPT-4o-mini GPT-4o Mistral-7B-Instruct-v0.3

3.1 Training setup

As a benchmark for all tests, we resort to the human quality evaluations corresponding to the science news articles we assess (N = 240). Both articles and reports have been published by the Medien-Doktor project, which offers the largest publicly available repository of medicine-related news reviews in a German context. The quality reports are consolidated summaries based on two independent evaluations for each article. For each article evaluation, the Medien-Doktor commissions two science journalism experts ("peers") specialised in the relevant field. These experts independently assess the texts based on the criteria presented above and then confer to agree on a consolidated overall evaluation. The evaluations include both a label (criterion *fulfilled*, *not fulfilled*, *not applicable*) and a brief

explanation of the decision taken for each criterion. As it is also the case in other peer review processes, this approach inherently introduces some level of disagreement between individual expert evaluations and the final summary ratings, providing a benchmark for reasonable evaluation discrepancies. For each article, the expert report that shows higher agreement with the consolidated summary is referred to as *human_high* in the following, while the remaining report is referred to as *human_low*. The terms "high" and "low" refer exclusively to the level of agreement with the consolidated expert summary and do not indicate differences in assessment quality. This joint report is treated as gold standard as it reflects the consensus reached by the editor between two reviewers. This procedure is illustrated in Figure 1.

Figure 1. Illustration of the evaluation procedure used in the Medien-Doktor project



Notes. Each article is independently reviewed by two domain-specialized science journalism experts, who produce individual quality reports based on the ten predefined criteria (*quality reports 1* & *2*). The reviewers then discuss their assessments and agree on a consolidated *final report*, which serves as the gold standard in our study. For each article, the individual report that aligns more closely with the final consensus is labelled human_high (in this example: Reviewer 1), while the one showing a greater deviation is labelled human_low (Reviewer 2).

3.2 Naive Bayes models

For our Naive Bayes approach, we train ten separate models, each specialised in predicting the assessment of one of the quality criteria by returning a single label Ɛ {fulfilled, not fulfilled}. This one-model-per-criterion setup is necessary since each quality criterion represents an independent binary classification task, and NB does not jointly model multiple labels in a single classifier. The models are

trained on a document feature matrix representing our training texts. We include only case-sensitive features (unigrams) that occur at least four times in the whole training data and in at least four documents. A small set of stop words (mainly pronouns and articles) and numbers is removed. This yields a medium-sized vocabulary that captures general lexical regularities without overfitting to idiosyncratic expressions in single articles. The final configuration was chosen based on a systematic comparison of 128 NB variants with different parameter combinations and n-gram settings based on our training data.

3.3 Random Forest classifier

In addition to this generic bag-of-words baseline, we train ten Random Forests, one for each of the ten quality criteria. In contrast to the NB models, the RFs operate on a compact, expert-informed feature set: For every criterion, the authors compiled a list of domain-specific signal words and regular expressions based on the established quality criteria (see Table 1) and their domain expertise in science journalism (e.g., terms referring to study design for "evidence" or to adverse effects for "risks"; see S1 b.). For each article and each criterion, we count how often these criterion-specific patterns occur and use these counts as the only features in the corresponding RF. Thus, each RF is trained on a small, criterion-tailored feature vector whose dimensionality equals the number of signal words defined for that criterion. This design deliberately differs from the NB setup: While NB serves as a classical, broad bag-of-words baseline, the RF models encode expert knowledge in a focused "bag-of-cues" representation, allowing us to test whether curated signal words alone can approximate human quality ratings. In the following, we refer to NB as well as the RF approaches as bag-of-words or "BOW" models in the sense that they disregard word order and rely exclusively on token or pattern counts.

3.4 Zero-Shot LLMs

When using an LLM "zero-shot", we give each model, one by one, an article as input, accompanied by the ten quality criteria presented earlier. All models are prompted to behave like a "rigorous scientific reviewer of health journalism articles", thereby following the well-established practice of assigning a specific role to an LLM to enhance its performance (Salewski et al., 2024; Shanahan et al., 2023). In each prompt, we include all criteria simultaneously along with the article, requiring the model to respond to all criteria sequentially as *fulfilled*, *not fulfilled* or *not applicable*. We also experimented with different prompting strategies, the differences in performance, however, remained minor. Specifically, our results showed that instructing the models to generate a full textual review before assigning labels slightly reduced accuracy and did not provide additional analytical value in the zero-shot setting (see S1 c. for further details on the prompting). In the following, when referring to a language model that is used in this vein, we use the term "zero-shot model".

3.5 Fine-tuned LLMs

Our "ft"-models are trained to generate standardised quality reports in natural language in the style of a human-written review, as the ones provided by the Medien-Doktor project (see S1 d. for an illustration). Specifically, the models process a raw document as input and output a detailed report consisting of ten paragraphs, each beginning with a brief assessment and concluding with a label ∈ [fulfilled, not fulfilled, not applicable]. The idea behind this approach is that, during training, the models internalise how the criteria are applied in context by "reading" expert evaluations and producing their own versions, with the fine-tuning process penalising deviations from the human assessments. This is a notable difference from the zero-shot LLMs, that apply the criteria based solely on fixed definitions they receive as part of the prompt.

We train all models on the same 212 documents of the human reviewers that were sampled randomly from the Medien-Doktor database. The remaining 28 documents serve as test data. The sampling ensures that the test set includes a representative distribution of article types and quality levels. Importantly, none of the models has access to the test articles during training, and no overlapping or near-duplicate content exists between the sets. The findings presented below are based on the models' (and humans') quality evaluations of the 28 test documents. For the evaluation, all predictions, including the LLM outputs and the expert classifications, are converted into a binary format by treating all "not applicable" labels as "not fulfilled" (see S1 a., supplementary material on OSF, for further details). Since each report evaluates the same ten criteria, the total number of data points for analysis is 280 (10 criteria × 28 documents).

To contextualise the dataset, Table 3 summarises the distribution of fulfilled and non-fulfilled labels across the 28 test documents. In addition, to give readers a sense of the types of articles analysed, four randomly selected headlines (English translations) are listed below (Table 4).

Table 3. Overview of label distribution across all ten criteria (C1–C10) in the test set. Values indicate the proportion of articles for which a given criterion was rated as fulfilled or not fulfilled

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
fulfilled	0.43	0.54	0.39	0.36	0.68	0.68	0.29	0.79	0.50	0.71
not fulfilled	0.57	0.46	0.61	0.64	0.32	0.32	0.71	0.21	0.50	0.29

Notes. C1 = Adequately explains benefits; C2 = Adequately explains potential harms; C3 = Reviews the study methodology or quality of evidence; C4 = Seeks independent sources and discloses conflicts of interest; C5 = Does not rely solely on a news release; C6 = Establishes the true novelty of the idea; C7 = Compares the new idea with existing alternatives; C8 = Establishes availability of the product or procedure; C9 = Adequately discusses costs; C10 = Avoids disease mongering.

Table 4. Excerpt of the dataset to be evaluated

Source	Date	**Title** (translated by the authors)
RND	2019-06-04	New coffee study: Up to 25 cups a day are safe
Der Standard	2020-02-17	Pregnancy: Overweight due to parabens in cosmetics
T-Online	2019-09-30	Cheese protects blood vessels from damage caused by excessive salt intake
Tagesspiegel	2020-08-04	Does Russia already have the coronavirus vaccine?

4. Results

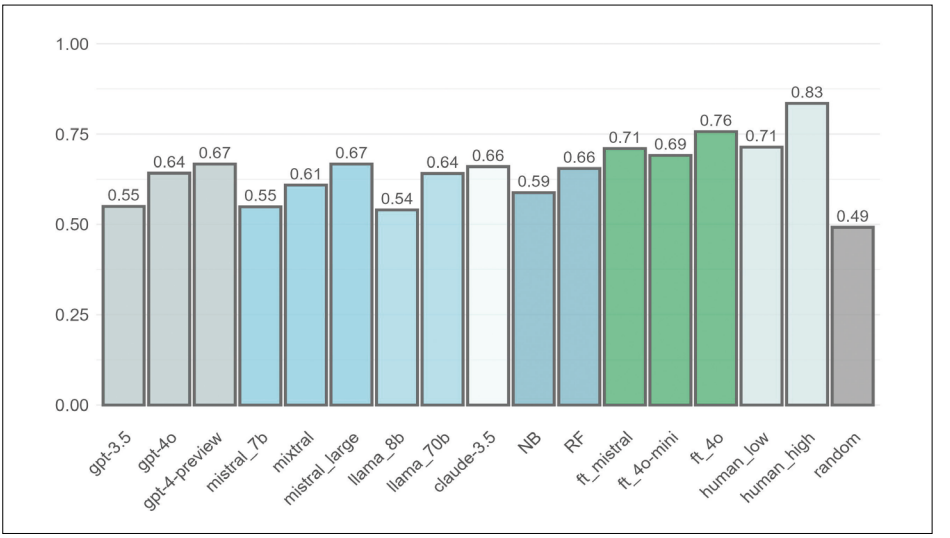
Based on the examined case of medical reporting, our findings suggest that ML models may offer useful support for assessing the quality of journalistic content. Particularly, fine-tuned LLMs appear to hold promise in assisting journalists with this task. However, some important differences remain, which may be relevant for the design of potential future assistance systems.

4.1 Overall model performances

Many tested models achieve reasonable accuracy rates in aligning their criterion evaluations with the consolidated expert assessments. For example, the best zero-shot LLMs, *GPT-4-1106-Preview* and *Mistral Large*, correctly predict 187 out of 280 labels, achieving accuracy rates of 67%, which is four percentage points below the values of the "*human_low* reports" (198 labels). However, since class imbalances affect our dataset, accuracy rates alone are insufficient to evaluate performance. Therefore, we additionally assess precision, recall, and macro-averaged F1-scores for both labels ("fulfilled" and "not fulfilled"). These numbers confirm that larger zero-shot LLMs come close to human expert results, and significantly outperform smaller and older models (e.g., *GPT-3.5-Turbo*, *Mistral 7B*, *Llama 8B*). Noteworthy, the RF classifier is among the best-performing models, closely matching the accuracy of the top-performing zero-shot LLMs (Figure 2). As described earlier, these rates should not be interpreted against certain reliability rates of a coding process (e.g., intercoder reliability) but against the background of current agreement rates in a peer-review process, where rather moderate levels of agreement among reviewers in similar contexts are common (Bornmann et al., 2010; Burla et al., 2008; Reinhart, 2009).

The highest performance is achieved by the fine-tuned LLMs, all of which reach F1-scores of 0.69 or higher and even surpass our "human_low" baseline (0.71). When comparing the zero-shot *Mistral 7B* to its fine-tuned counterpart, the performance gains from fine-tuning become particularly evident (0.55 vs. 0.71 F1 score).

Figure 2. Macro-averaged F1 scores achieved by all tested models



Notes. The different colour sets indicate the different model families: *OpenAI* (grey), *Mistral* (blue), *Llama* (light blue), *Claude* (white), BOW models (dark blue), fine-tuned LLMs (green; "ft_" indicates fine-tuning), human experts (light grey). For comparison, the figure also displays the performances of a hypothetical model that would sample randomly from both labels ("random"). Values reflect agreement with the subjective evaluations by (other) experts rather than with an objective ground truth. As in other peer-review processes, moderate levels of interrater agreement are expected.

Beyond quantitative performance, a heuristic examination of the quality reports generated by the fine-tuned models provides further insights into the models' reasoning processes and their alignment with human evaluations. We find that fine-tuning indeed taught the models to focus on similar aspects as human experts, such as prioritizing absolute over relative numbers when assessing an article's description of the benefits (criterion 1) or risks (criterion 2) of a treatment (see S4, supplementary material on OSF for an example). For the smaller two ft-models, the results are not as clear-cut. While performing well in quantitative terms, a heuristic analysis of the generated reports reveals that the smaller models partly learned to overfit the training data. In an illustrative example, the *fine-tuned Mistral* model learned to always classify criterion 10 as fulfilled, stating that "[topic of article] is not exaggerated". This clearly shows the importance of model selection and parameter tuning, a challenge we want to tackle further in future research.

These data let us answer research questions 1–2 as follows: Regarding RQ1, our results demonstrate that automated quality assessment of scientific news articles is indeed feasible, with fine-tuned LLMs achieving F1 scores of up to 0.76 – a performance that surpasses that of human experts. Given both the performance metrics and the substantive quality of the model-generated reports, the *fine-tuned GPT-4o* appears capable of producing assessments as reliable as those written by the lower-performing half of the human experts.

As for RQ2, concerning the minimum technical requirements needed for reliable quality assessments, our findings show that while simple NB models and older LLMs like *GPT-3.5-Turbo* perform only marginally better than random guessing, an RF classifier can match the performance of more complex LLMs. We conclude that while using more recent *ChatGPT* variants with detailed prompts (which is a common procedure in many newsrooms) provides a convenient approach with reasonable accuracy, similar outcomes can be achieved with a simpler RF model. This confirms earlier findings by Liu et al. (2025) and underlines the importance of not focusing on popular LLMs like *ChatGPT* alone (which is often the case in public discourse), but also considering simpler, more interpretable models where appropriate. For optimal performance, however, fine-tuning an LLM appears essential.

4.2 Criterion-specific performance

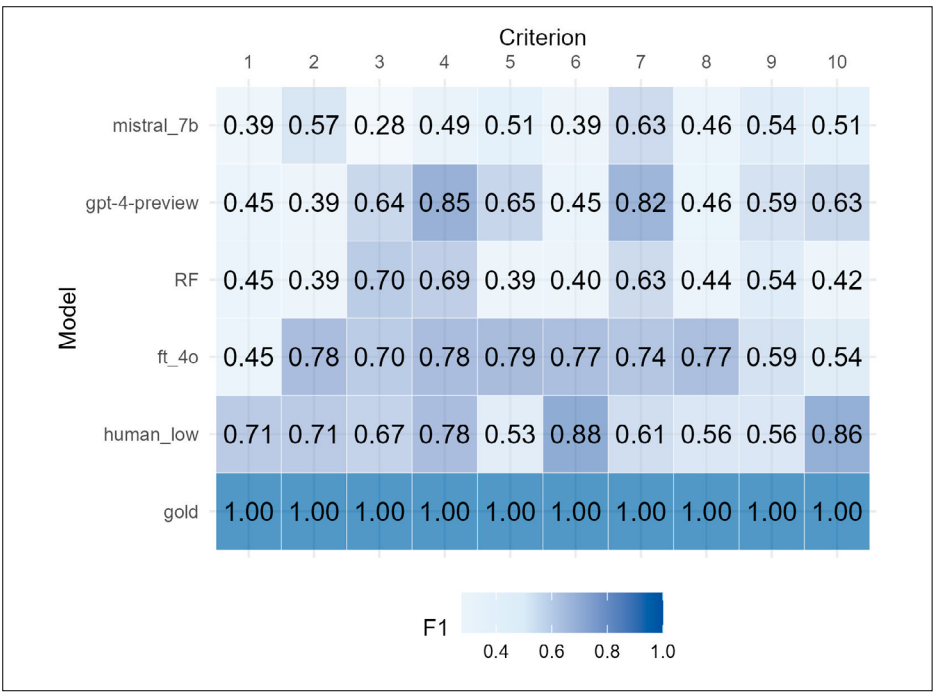
To better understand the strengths and weaknesses of the tested models, we evaluate their performance on each of the ten criteria individually (S3 b.). The results show that, while the prediction of certain criteria appears straightforward, multiple criteria remarkably challenge the tested models (Figure 3). Especially the criteria 1 (potential benefits) and 10 (neutrality/avoiding disease mongering) prove difficult to classify, with all models frequently diverging from human expert conclusions (min. 0.2 difference in avg. F1 score, respectively). While the fine-tuned language models achieve comparatively strong results, they still struggle with these two criteria.

A likely explanation for this lies in the underlying training data distribution. For certain criteria, especially criterion 10, most gold-standard labels are "fulfilled" (20 out of 28). As a result, the ft-models appear to have learned this dominant pattern well, yielding high recall and precision for positive predictions, but struggle to reliably identify rare "not fulfilled" cases. This issue might be mitigated in future through targeted adjustments to the training data, such as data augmentation (see Discussion). However, such adjustments will not be a universal fix. For example, criterion 9 (costs) also poses difficulties, yet here even human experts show relatively poor agreement, suggesting that some criteria may be inherently harder to assess. In the context of journalistic practice, these findings once again demonstrate the importance of understanding a model's "behaviour" when relying on its outputs. In our case, this would mean that predictions for criteria 1 and 10 would require special attention, with human judgement assuming a particularly prominent role in their evaluation.

In sum, regarding RQ3, our findings reveal substantial variation in prediction accuracy across quality criteria. While some criteria are predicted with reasonable recall and precision, multiple consistently prove challenging for both BOW and the zero-shot LLMs. Only the fine-tuned LLMs, especially the *fine-tuned GPT-4o* model, manage to reliably predict eight out of ten criteria, demonstrating the benefits of training on complete assessment reports. However, the persistent difficulty with specific criteria suggests that even fine-tuned models may struggle to gene-

ralize across all dimensions of article quality, pointing to structural limitations that warrant further investigation.

Figure 3. Criterion-wise model performance



Notes. The heatmap shows macro-averaged F1 scores for each of the ten quality criteria (columns) across a selection of models (rows). Included are: *Mistral 7B* (representing smaller open-source LLMs), *GPT-4-1106-Preview* (as the best-performing zero-shot LLM), Random Forest (RF; best-performing bag-of-words model), and a *fine-tuned GPT-4o* model (ft_4o). Human performance is represented by the less-aligned reports (*human_low*) and the expert consensus (*gold*). The figure highlights that while certain criteria are reliably predicted by most models, others, such as criteria 1 and 10, remain challenging even for fine-tuned LLMs.

4.3 Quality-based article selection

In practical applications, the overall assessment of an article is arguably more important than a model's accuracy on individual criteria. To address this, we also evaluate the models' ability to perform a "gatekeeping" function by distinguishing between high- and low-quality articles, based on a document's number of fulfilled criteria. To this end, we categorise each article based on its expert rating: Articles with fewer than five fulfilled criteria are considered "low quality," those with only one or two fulfilled criteria are classified as "bad". All other articles are classified as "pass", i.e., articles that fulfil five or more of the ten criteria.

Our experiments show that all top-performing models (RF, *gpt-4-preview*, *ft_4o*) successfully classify all the worst articles (n = 5) as either low-quality or

bad (see S5 for an illustration). Given the small number of worst articles, this result should be interpreted as illustrative rather than statistically conclusive. Overall, the best-performing model (*fine-tuned GPT-4o*) correctly classifies 21 out of 28 articles, achieving higher agreement with the consolidated expert assessment than the less-aligned expert reports (*human_low*, 19/28). The other models, in contrast, show substantial variation in their class prediction capabilities, with, for example, the zero-shot *GPT-4o* model misclassifying two of the best articles as low-quality. This illustrates two things: First, the benefits of fine-tuning an LLM on expert data, and second, the inherent complexity and subjectivity of the task, as even human experts do not always fully agree on which articles should be considered low quality. Table 5 summarises the results in the form of a confusion matrix (recall, precision, F1).

Table 5. Macro-averaged precision, recall, and F1-scores across the three quality classes (bad, low quality, pass) for selected models and the human expert baseline

	precision	recall	F1
gpt-4-preview	0.46	0.44	0.43
RF	0.42	0.53	0.47
ft_4o	0.68	0.73	0.69
human_low	0.57	0.56	0.57

Considering this, we conclude (RQ4) that various types of models can provide meaningful assistance in editorial workflows, particularly during filtering stages. Again, the *fine-tuned GPT-4o* model delivered the most reliable results. However, none of these models should be seen as a substitute for expert review. They should instead be understood as complementary tools that support, rather than replace, human judgement.

5. Discussion

The results of our study indicate that the evaluation and potential adoption of AI models for content evaluation beyond fact-checking is a promising direction for both journalism and science communication research. We show that clearly defined quality criteria, particularly in a scientific domain, can be used for ML-based content-evaluation of articles, given these models are either properly trained or large enough to be able to perform an analysis zero-shot. For incoming texts such as press releases from scientific institutions, agency materials or pieces from freelancers, a reliably performing model could provide initial quality checks and identify articles that appear most promising for further (human) investigation. Before the final release of journalistic texts, it could furthermore serve as an outbound control, ensuring that content meets established quality criteria of a certain media house or institution. However, among the tested models, only the fine-tuned LLMs, particularly the *fine-tuned GPT-4o*, reached a level of performance that makes them credible candidates for potential application in newsroom environ-

ments. Smaller and older models such as *GPT-3.5* performed significantly worse, especially compared to traditional ML models. This finding complements the results presented by Liu et al. (2025), where *GPT-3.5-Turbo* also showed limited accuracy in a zero-shot setting and underperformed relative to supervised baselines. Our results indicate that this outcome is largely attributable to the model's limited capabilities rather than a fundamental inability of LLMs to evaluate journalistic quality. More recent and larger models substantially outperform *GPT-3.5* in our tests and reach performance levels comparable to our Random Forest classifier. Crucially, we find that fine-tuning LLMs on full-text expert reports leads to further improvements, enabling the best-performing model to achieve higher agreement with the consolidated expert assessments than individual human reviewers. Furthermore, it comes with the benefit of providing users with a detailed report that discusses an article's quality regarding each of the ten quality criteria, which can be a valuable support in journalistic workflows.

At the same time, when it comes to the real-world adaptability of our framework, we acknowledge that fine-tuning a language model is not a trivial task, and that newsrooms might face challenges setting up their own quality-evaluation models. While ML expertise may play a role, the primary obstacle probably remains the lack of high-quality training data, which critically constrains a model's ability to learn nuanced quality distinctions (see below). With this in mind, there is a need for further investigation into how to further push the zero-shot performance of LLMs. Comprehensive analyses of where and why discrepancies between model and human evaluations occur could lead to more targeted improvements in model prompting. We are confident that through extensive prompt engineering, the performance of the zero-shot models analysed here can be further improved. However, whenever sufficiently high-quality training data is available, fine-tuning should be the preferred approach.

For any real-world application, it remains essential to emphasise that even the most capable model cannot and should not replace journalistic gatekeeping responsibilities. While models can assist in initial screening, they cannot and should not fully replace human oversight and critical judgement, which remain core pillars of the journalistic profession (Gutiérrez Lopez et al., 2023; Lin & Lewis, 2022). This is especially true for nuanced editorial decisions, when model outputs are ambiguous or when a model has already demonstrated systematic limitations in certain settings. For example, our results have shown that even a fine-tuned LLM that achieves higher agreement with the consolidated expert assessments than individual human reviewers in many tests shows systematic biases in predicting certain quality criteria. Encouragingly, recent studies confirm that most journalists, even when actively using AI tools to enhance efficiency, maintain a critical stance, insisting on verifying and editing AI outputs to ensure alignment with journalistic standards (Wu, 2024).

6. Limitations

Despite the promising results of our study, several limitations must be acknowledged. Firstly, our analysis focuses exclusively on the example of articles related to medical science and relies on the existence and testability of predefined quality criteria. Thus, the adaptability of our approach to other areas of journalism remains to be explored. We are, however, confident that given the existence of clear quality standards, our approach could be effectively applied in other fields. In particular, the utility of these models for economic or environmental reporting (where corresponding quality criteria have already been developed (Rögener & Wormer, 2017)) seems worth exploring in future research. Secondly, although we tested a variety of models and prompt techniques, the generalisability of our findings could be further strengthened by incorporating other popular (large language) models. In particular, resorting to more recent "thinking models" like *DeepSeek R1* or OpenAI's *o3* model could be promising and should be evaluated in future research. The same applies to the potential of including tool-use or internet search capabilities in the LLM setup. Both extensions could allow LLMs to access external information, potentially enhancing their performance on tasks requiring up-to-date or context-specific knowledge. Since long-term independence from proprietary model providers such as OpenAI or Alphabet might be a relevant strategic consideration for many news organisations, open-source models such as *Mistral*, *Gemma 3* (Google) or *Phi 4* (Microsoft) deserve particular attention in this context. Many of these models can now be re-trained or adapted with comparatively modest computational resources, including setups based on a cloud platform such as Google Colab or Kaggle, indicating that, from a technical standpoint, the prerequisites for adopting and customising open-source solutions are already quite favourable. Finally, our findings represent a snapshot in time. Given the rapid pace of development in large language models, future architectures or model versions may exhibit different performance characteristics, even on comparable tasks.

A further critical limitation is the restricted size of our dataset. We are well aware that incomplete or limited training data contributes to a lack of robustness and prediction accuracy. Moreover, as with any fine-tuning on relatively small datasets, overfitting remains a potential concern – although our best-performing models showed no indications of overfitting on the validation or unseen test data (a brief analysis of the training dynamics is provided in supplementary material S1 d on OSF). However, the limited availability of publicly accessible expert reviews on science news articles, such as those presented here, constrains our ability to scale up the training data. One potential strategy to mitigate data scarcity is data augmentation (e.g., Feng et al., 2021; Wei & Zou, 2019). However, implementing such techniques in our context is challenging: Both the original articles and their corresponding quality assessments would need to be adapted in a highly controlled and semantically consistent manner. As an alternative or complementary strategy, collaborations with media organizations could offer a more sustainable path forward. By systematically logging editorial decisions over time (e.g., which texts are selected, edited, or discarded, and why so), these organizations could generate domain-specific training data that reflects real-world workflows and decisions.

7. Conclusions

Budget constraints in media organisations, coupled with an increasing need for efficient quality assessment amid growing information flows, have intensified the demand for tools that can support rapid and reliable decision-making. This is especially true in fields such as science reporting, where specific editorial knowledge is often rare. Addressing this need, our study combines theoretical insights from quality research with practical model evaluations, testing a variety of AI approaches for their potential to automate criteria-based quality assessment of news articles, particularly in the domain of medical science reporting. This approach moves beyond the established, albeit important, practice of fact-checking specific claims (Hameleers, 2019; Nyhan et al., 2020) towards a more holistic evaluation of journalistic content based on predefined criteria for the overall quality of a given contribution.

Rather than relying solely on the "black box" of LLMs, we conducted a comparative analysis to evaluate the capabilities of both LLMs and more transparent methods in terms of their ability to evaluate an article's quality. By testing a range of models against established quality criteria for medical journalism, we found that current AI technologies can deliver results that are, in many cases, comparable to human expert evaluations. Both advanced LLMs, such as *GPT-4o*, and simpler models like a Random Forest classifier achieved accuracy levels approaching those of human evaluators but showed imbalances in their prediction capabilities of individual quality criteria. Notably, the fine-tuned LLMs employed in this study showed a significant advantage over their zero-shot counterparts. A *fine-tuned GPT-4o* model consistently aligned more closely with – and even surpassed – human expert evaluations. Nevertheless, at least two quality criteria proved challenging also for the fine-tuned models, underscoring the complexity and nuanced understanding needed for robust quality assessment in journalism. Importantly, the fine-tuned models also demonstrated strong capabilities in detecting low-quality articles, suggesting practical applications as assistive tools in editorial workflows, particularly during early screening stages. While our study focused on medical journalism, we are confident that our approach can be extended to other fields where clear quality standards exist or can be developed, tailored to the needs of specific organisations or aligned with theories of media performance. Ultimately, the models tested here should always be seen as augmenting, rather than replacing, human judgment.

Generative AI declaration

During the preparation of this work, the authors used DeepL and, in rare cases, ChatGPT to refine the linguistic quality and improve the readability of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Supplementary material

Supplementary material for this article is available on the Open Science Framework (OSF): https://osf.io/6yk82.

References

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School (HKS) Misinformation Review*, *3*, 1–15.

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21). https://doi.org/10.1073/pnas.2314021121

Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital journalism*, *7*(2), 206–229. https://doi.org/10.1080/21670811.2018.1521292

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE*, *5*(12). https://doi.org/10.1371/journal.pone.0014331

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bucchi, M., & Mazzolini, R. G. (2003). Big science, little news: Science coverage in the Italian daily press, 1946–1997. *Public Understanding of Science*, *12*(1), 7–24. https://doi.org/10.1177/0963662503012001413

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research*, *57*(2), 113–117. https://doi.org/10.1097/01.NNR.0000313482.33917.7d

Cacciatore, M. A. (2021). Misinformation and public opinion of science and health: Approaches, findings, and future directions. *Proceedings of the National Academy of Sciences*, *118*(15). https://doi.org/10.1073/pnas.1912437117

Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, *3*(3), 416–431. https://doi.org/10.1080/21670811.2014.976412

Caswell, D., & Dörr, K. (2018). Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice*, *12*(4), 477–496. https://doi.org/10.1080/17512786.2017.1320773

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*, 1–45. https://doi.org/10.1145/3641289

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv*, arXiv:2403.04132. https://doi.org/10.48550/arXiv.2403.04132

Choi, S., Shin, H., & Kang, S.-S. (2021). Predicting audience-rated news quality: Using survey, text mining, and neural network methods. *Digital Journalism*, *9*(1), 84–105. https://doi.org/10.1080/21670811.2020.1842777

Clark, F., & Illman, D. L. (2006). A longitudinal study of the New York Times science times section. *Science Communication*, *27*(4), 496–513. https://doi.org/10.1177/1075547006288010

Cohen, S., Hamilton, J. T., & Turner, F. (2011). Computational journalism. *Communications of the ACM*, *54*(10), 66–71. https://doi.org/10.1145/2001269.2001288

Cordero, J. M., Henn, T., Holtel, F., Sánchez Gómez, J. Á., Arenas, D., Šipka, A., & Vollmer, S. (2025). Developing effective and value-aligned AI tools for journalists: 12 critical questions to reflect upon. *Journalism Practice*, 1–20. https://doi.org/10.1080/17512786.2025.2465894

Davies, N. (2008). *Flat earth news: An award-winning reporter exposes falsehood, distortion and propaganda in the global media*. Chatto & Windus.

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Harvard University Press. https://doi.org/10.4159/9780674239302

Diakopoulos, N., Cools, H., Helberger, N., Li, C., Kung, E., & Rinehart, A. (2024). *Generative AI in journalism: The evolution of newswork and ethics in a generative information ecosystem*. Associated Press. https://doi.org/10.13140/RG.2.2.31540.05765

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. *Digital Journalism*, *4*(6), 700–722. https://doi.org/10.1080/21670811.2015.1096748

Fähnrich, B., Weitkamp, E., & Kupper, J. F. (2023). Exploring 'quality' in science communication online: Expert thoughts on how to assess and promote science communication quality in digital media contexts. *Public Understanding of Science*, *32*(5), 605–621. https://doi.org/10.1177/09636625221148054

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*. https://doi.org/10.18653/v1/2021.findings-acl.84

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30). https://doi.org/10.1073/pnas.2305016120

Gladney, G. A. (1996). How editors and readers rank and rate the importance of eighteen traditional standards of newspaper excellence. *Journalism & Mass Communication Quarterly*, *73*(2), 319–331. https://doi.org/10.1177/107769909607300204

Gutiérrez Lopez, M., Porlezza, C., Cooper, G., Makri, S., MacFarlane, A., & Missaoui, S. (2023). A question of design: Strategies for embedding AI-driven tools into journalistic work routines. *Digital Journalism*, *11*(3), 484–503. https://doi.org/10.1080/21670811.2022.2043759

Hameleers, M. (2019). Susceptibility to mis- and disinformation and the effectiveness of fact-checkers: Can misinformation be effectively combated? *SCM Studies in Communication and Media*, *8*(4), 523–546. https://doi.org/10.5771/2192-4007-2019-4-523

Helberger, N., van Drunen, M., Moeller, J., Vrijenhoek, S., & Eskens, S. (2022). Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals. *Digital Journalism*, *10*(10), 1605–1626. https://doi.org/10.1080/21670811.2022.2152195

Hermida, A., & Simon, F. M. (2025). AI in the newsroom: Lessons from the adoption of The Globe and Mail's Sophi. *Journalism Practice*, *19*(10), 1–18. https://doi.org/10.1080/17512786.2025.2471781

Himma-Kadakas, M., & Ojamets, I. (2022). Debunking false information: Investigating journalists' fact-checking skills. *Digital Journalism*, *10*(5), 866–887. https://doi.org/10.1080/21670811.2022.2043173

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Social Network Analysis and Mining*, *10*(1). https://doi.org/10.1007/s13278-020-00696-x

Jindal, H., Mangla, M., & Singh, G. (2023, June 16–17). *Fake news detection using machine learning*. International Conference on Recent Developments in Cyber Security (pp. 375–385), Greater Noida, India. https://doi.org/10.1016/j.procs.2025.09.048

Kim, B., Xiong, A., Lee, D., & Han, K. (2021). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PLoS ONE*, *16*(12). https://doi.org/10.1371/journal.pone.0260080

Kotenidis, E., & Veglis, A. (2021). Algorithmic journalism – current applications and future perspectives. *Journalism and Media*, *2*(2), 244–257. https://doi.org/10.3390/journalmedia2020014

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, *31*(2). https://doi.org/10.1002/ett.3767

Latar, N. L. (2018). *Robot journalism: Can human journalism survive?*. World Scientific. https://doi.org/10.1142/10913

Lin, B., & Lewis, S. C. (2022). The one thing journalistic AI just might do for democracy. *Digital Journalism*, *10*(10), 1627–1649. https://doi.org/10.1080/21670811.2022.2084131

Liu, X., He, L., Alanazi, E., Liu, E., Goss, A., & Gumireddy, L. (2025). Assessing the accuracy and explainability of using ChatGPT to evaluate the quality of health news. *BMC Public Health*, *25*(1). https://doi.org/10.1186/s12889-025-23206-0

Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., ..., & Shah, S. (2016). Reuters tracer: A large scale system of detecting & verifying real-time news events from Twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 207–216). Association for Computing Machinery. https://doi.org/10.1145/2983323.2983363

Løvlie, A. S., Waagstein, A., & Hyldgård, P. (2023). "How trustworthy is this research?" Designing a tool to help readers understand evidence and uncertainty in science journalism. *Digital Journalism*, *11*(3), 431–464. https://doi.org/10.1080/21670811.2023.2193344

LSE (London School of Economics and Political Science) (2025). *JournalismAI Case Studies*. Retrieved March 13, 2026, from https://airtable.com/appdaeDwFizD4RK0u/shrKhe7Js48HvBhmG/tblBcSZESOAuy5Q9A

Mahony, S., & Chen, Q. (2025). Concerns about the role of artificial intelligence in journalism, and media manipulation. *Journalism*, *26*(9). https://doi.org/10.1177/14648849241263293

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (pp. 41–48). AAAI Press.

McQuail, D. (1993). *Media performance: Mass communication and the public interest*. University of Toronto Press. https://doi.org/10.22230/cjc.1993v18n4a783

Meier, K. (2019). Quality in journalism. In T. P. Vos, F. Hanusch, D. Dimitrakopoulou, M. Geertsema-Sligh, & A. Sehl (Eds.), *The International Encyclopedia of Journalism Studies*. Wiley. https://doi.org/10.1002/9781118841570.iejs0041

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241231468

Møller, L. A., van Dalen, A., & Skovsgaard, M. (2025). A little of that human touch: How regular journalists redefine their expertise in the face of artificial intelligence. *Journalism Studies*, *26*(1), 84–100. https://doi.org/10.1080/1461670X.2024.2412212

Noain Sánchez, A. (2022). Addressing the impact of artificial intelligence on journalism: The perception of experts, journalists and academics. *Communication & Society*, *35*(3), 105–121. https://doi.org/10.15581/003.35.3.105-121

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*(3), 939–960. https://psycnet.apa.org/doi/10.1007/s11109-019-09528-x

Oxman, A. D., Guyatt, G. H., Cook, D. J., Jaeschke, R., Heddle, N., & Keller, J. (1993). An index of scientific quality for health reports in the lay press. *Journal of Clinical Epidemiology*, *46*, 987–1001. https://doi.org/10.1016/0895-4356(93)90166-x

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? *arXiv*, arXiv:2302.06476. https://doi.org/10.18653/v1/2023.emnlp-main.85

Reich, Z., & Godler, Y. (2014). A time of uncertainty: The effects of reporters' time schedule on their work. *Journalism Studies*, *15*(5), 607–618. https://doi.org/10.1080/1461670X.2014.882484

Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics*, *81*(3), 789–809. https://doi.org/10.1007/s11192-008-2220-7

Rieger, J., Yanchenko, K., Ruckdeschel, M., von Nordheim, G., Kleinen-von Königslöw, K., & Wiedemann, G. (2024). Few-shot learning for automated content analysis: Efficient coding of arguments and claims in the debate on arms deliveries to Ukraine. *SCM Studies in Communication and Media*, *13*(1), 72–100. https://doi.org/10.5771/2192-4007-2024-1-72

Rögener, W., & Wormer, H. (2017). Defining criteria for good environmental journalism and testing their applicability: An environmental news review as a first step to more evidence based environmental science reporting. *Public Understanding of Science*, *26*(4), 418–433. https://doi.org/10.1177/0963662515597195

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., & Akata, Z. (2024). In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, *36*, 72044–72057. https://doi.org/10.48550/arXiv.2305.14930

Schäfer, M. S. (2023). The notorious GPT: Science communication in the age of artificial intelligence. *JCOM: Journal of Science Communication*, *22*. https://doi.org/10.22323/2.22020402

Schäfer, M. S., Kremer, B., Mede, N. G., & Fischer, L. (2024). Trust in science, trust in ChatGPT? How Germans think about generative AI as a source in science communication. *Journal of Science Communication*, *23*(9). https://doi.org/10.22323/2.23090204

Schwitzer, G. (2008). How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. *PLoS Medicine*, *5*(5). https://doi.org/10.1371/journal.pmed.0050095

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Silva Luna, D., Broer, I., Bilandzic, H., Taddicken, M., Schuller, B. W., & Bürger, M. (2025). Quality in science communication with communicative artificial intelligence: A principle-based framework. *Public Understanding of Science*, *34*(8). https://doi.org/10.1177/09636625251328854

Simon, F. M. (2024). *Artificial intelligence in the news: How AI retools, rationalizes, and reshapes journalism and the public arena (tech. rep.)*. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/ncm5-3v06

Smith, D. E., Wilson, A. J., & Henry, D. A. (2005). Monitoring the quality of medical news reporting: Early experience with media doctor. *Medical Journal of Australia*, *183*(4), 190–193. https://doi.org/10.5694/j.1326-5377.2005.tb06992.x

Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289

Summ, A., & Volpers, A.-M. (2016). What's science? Where's science? Science journalism in German print media. *Public Understanding of Science*, *25*(7), 775–790. https://doi.org/10.1177/0963662515583419

Tankard, J. W., Jr., & Ryan, M. (1974). News source perceptions of accuracy of science coverage. *Journalism Quarterly*, *51*(2), 219–225. https://doi.org/10.1177/107769907405100204

Tay, L. Q., Lewandowsky, S., Hurlstone, M. J., Kurz, T., & Ecker, U. K. (2024). Thinking clearly about misinformation. *Communications Psychology*, *2*(1). https://doi.org/10.1038/s44271-023-00054-5

Thurman, N. (2019). Computational journalism. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (2nd ed., pp. 180–195). Routledge.

Van Dalen, A. (2012). The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice*, *6*(5–6), 648–658. https://doi.org/10.1080/17512786.2012.667268

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*. https://doi.org/10.18653/v1/D19-1670

West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, *118*(15). https://doi.org/10.1073/pnas.1912444117

Wilson, A., Bonevski, B., Jones, A., & Henry, D. (2009). Media reporting of health interventions: Signs of improvement, but major problems persist. *PLoS ONE*, *4*(3). https://doi.org/10.1371/journal.pone.0004831

Wolf, C. (2024). Journalismus und künstliche Intelligenz aus kommunikationswissenschaftlicher Perspektive: Chancen und Herausforderungen [Journalism and artificial intelligence from a communication science perspective: Opportunities and challenges]. In G. Hooffacker, W. Kenntemich, & U. Kulisch (Eds.), *Neue Plattformen – neue Öffentlichkeiten: KI, Krisen und Journalismus* (pp. 9–29). Springer Fachmedien. https://doi.org/10.1007/978-3-658-44659-8_2

Wormer, H. (2011). Improving health care journalism. In G. Gigerenzer & J. A. Muir Gray (Eds.), *Better doctors, better patients, better decisions: Envisioning Health Care 2020* (pp. 169–188). MIT Press. https://doi.org/10.7551/mitpress/9143.003.0016

Wu, S. (2024). Journalists as individual users of artificial intelligence: Examining journalists' "value-motivated use" of ChatGPT and other AI tools within and without the newsroom. *Journalism*. https://doi.org/10.1177/14648849241303047

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv*, arXiv:2303.17564. https://doi.org/10.48550/arXiv.2303.17564

Zhang, H. (2004). The optimality of naive Bayes. In V. Barr & Z. Markov (Eds.), *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (pp. 562–567). AAAI Press.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*, 237–291. https://doi.org/10.1162/coli_a_00502