Fact-Checking Sustainability Objectives Using Multimodal Retrieval-Augmented Generation

Mohammad Mahdavi
Gisma University of Applied Sciences
Potsdam, Germany
mohammad.mahdavi@gisma.com

Amirhosein Farahmand
University of Tehran
Tehran, Iran
amir.farahmand@ut.ac.ir

Abolfazl Nadi
University of Tehran
Tehran, Iran
a.nadi@ut.ac.ir

***Abstract*—Sustainable development is now an important consideration for various stakeholders, including customers, investors, auditors, and policymakers. Although companies publish a range of reports to disclose their sustainability information, analyzing and fact-checking their claims is challenging due to the large volume and heterogeneity of the data. In this paper, we present our novel sustainability objective fact-checking system, which retrieves relevant multimodal textual and visual evidence to generate verification reports. We systematically evaluate our system and demonstrate that it effectively supports users across a variety of real-world fact-checking scenarios.**

***Index Terms*—sustainability, greenwashing, fact-checking, information retrieval, multimodal, retrieval-augmented generation**

I. Introduction

Sustainable development aims to meet present needs without compromising the ability of future generations to do the same [1]. To this end, the United Nations introduced Sustainable Development Goals (SDGs) in 2015 to tackle universal problems like climate change, poverty, and environmental issues [2]. Additionally, the Paris Agreement set climate targets for countries [3]. These sustainability milestones have driven companies and governments to be more mindful of their impact on the environment and society [4]. Companies strive to be perceived as eco-friendly in response to customers' demand for greener products and investors' preference for low-risk opportunities. Companies are publishing their environmental/social vision in various sustainability reports, such as Corporate Social Responsibility (CSR) and Environmental, Social, and Corporate Governance (ESG) [5]. In these reports, they demonstrate their commitment to aligning their profiles with the goal of being environmentally friendly, or in other words, "green" [6].

However, sustainability reporting has also revealed its dark side in recent years, with the spread of misleading information about the true impact of corporate environmental and social strategies [4], [7]. This phenomenon is commonly referred to as *greenwashing*, in which companies want to look greener in the eyes of the public than they really are [5]. A common example is when polluting companies provide false claims or information in their sustainability reports to appear more responsible [8]. This misinformation misleads consumers and stakeholders, eroding their trust in the entire green market [9].

Greenwashing detection aims at identifying such unsubstantiated environmental/social claims in sustainability reports of companies. This task consists of two general steps: identifying environmental/social claims of companies and then fact-checking them using all available evidence [10]. More specific sustainability claims, such as "we will reach net-zero carbon emissions by 2030," are referred to as *sustainability objectives* [5]. We have already developed the sustainability objective detection system, GoalSpotter, which identifies environmental and social claims in company reports [5]. The system automates objective detection by formulating it as a text classification task, labeling each text block in a sustainability report as either an objective or noise. [5].

The subsequent fact-checking step is more challenging. Once sustainability claims are identified in a report, relevant evidence must be gathered to verify them. Traditionally, domain experts manually search for supporting evidence across all of a company's sustainability reports, which is a time-consuming and non-scalable approach [10]. The fact-checking challenge becomes even greater when the evidence search involves a multimodal approach, requiring both textual and visual evidence to verify sustainability claims.

We propose and demonstrate a new sustainably objective fact-checking system using multimodal Retrieval-Augmented Generation (RAG). Specifically, we make the following contributions.

- We present a new benchmark framework for the multimodal fact-checking of sustainability objectives, incorporating sustainability reports, their underlying objectives, and corresponding textual and visual evidence.
- We design a novel fact-checking system for sustainability objectives that incorporates a multimodal RAG component to retrieve both textual and visual evidence. Our system is available online[1].
- We systematically evaluate our fact-checking system against a baseline and demonstrate its superiority across syntactic, semantic, and human evaluation measures.
- We demonstrate our sustainability-focused fact-checking system across various real-world scenarios, highlighting its unique features, effectiveness, and efficiency.

[1]https://github.com/m-mahdavi/sustainability-fact-checker

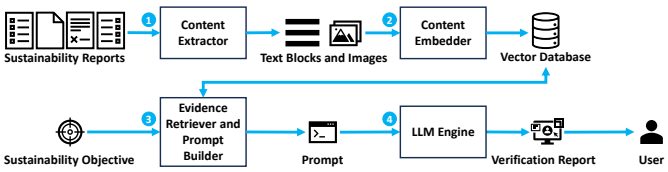

Fig. 1. The workflow of the system.

II. Multimodal Fact-Checking System

Problem formulation. Our goal is to support domain experts in sustainability objective fact-checking using historical reports of companies. Suppose $D = \{d_1, d_2, \ldots, d_{|D|}\}$ is the set of historical sustainability reports for a company. Let $O = \{o_1, o_2, \ldots, o_{|O|}\}$ be the set of sustainability objectives for the same company that have already been detected using objective detection systems, such as GoalSpotter [5]. For each sustainability objective o_i, the problem is to generate a verification report using the set of historical documents D to support the user in the fact-checking process. The verification report should contain a concise and comprehensive textual statement and relevant visuals that support or refute the given sustainability objective.

System overview. Verifying sustainability claims from company reports poses several technical challenges. First, companies publish a large number of lengthy reports over time, containing both textual and visual evidence. Extracting relevant information is difficult due to the heterogeneity and volume of the data. Raw text and image extraction produces a large unstructured data lake that cannot be efficiently searched. Second, retrieved evidence must be semantically relevant to the query, requiring a unified representation of both text and images. Finally, Large Language Models (LLMs), while powerful, are prone to bias and hallucination [11], [12]. Without careful grounding in retrieved evidence, LLMs may generate verification reports that are incoherent or non-factual.

To address these challenges, we design a fact-checking system that (i) semantically stores multimodal evidence from sustainability reports, (ii) retrieves relevant content based on user queries, and (iii) generates coherent verification reports grounded in retrieved evidence. Figure 1 illustrates the workflow of our sustainability objective verification system. Given a collection of sustainability reports from a company and a specific sustainability objective, the system generates a verification report to assist users in the fact-checking process. The system begins by parsing the company's sustainability reports to extract text blocks and images. It then embeds this content and stores it in a vector database. Using the provided sustainability objective, the system retrieves the most relevant text blocks and images to construct a prompt that incorporates the retrieved evidence. Finally, it uses an LLM to generate the verification report for the user.

Fact-checking procedure. Algorithm 1 provides the pseudocode. For each company, our system collects all sustainability reports $D = \{d_1, \ldots, d_{|D|}\}$. Each report d_i is preprocessed

Algorithm 1: SustainabilityObjectiveFactChecking(o, D)

Input: sustainability objective o, set of company reports D.
Output: verification report r with textual and visual evidence.

```
// -- The Offline Preprocessing Phase --
foreach report d_i ∈ D do
    B_i ← segment the text of report d_i into blocks;
    if segmentation is noisy then
        B_i ← refine text blocks B_i using an LLM;
    I_i ← extract images from report d_i;
    store embeddings of text blocks B_i and images I_i in the vector database;
// -- The Online Report Generation Phase --
q ← embed the sustainability objective o;
E ← retrieve top relevant text blocks/images from the database for query q;
E' ← ∅;
foreach retrieved text block b_i ∈ E do
    E' ← add retrieved neighboring text blocks within window size W;
r ← generate a verification report using evidence E and contexts E';
return r;
```

to extract textual and visual content. The text on each page is segmented into logical, self-contained blocks (e.g., headings, sentences, and paragraphs). When clean segmentation is difficult (e.g., in old PDF files), we leverage an open-source lightweight LLM (in our prototype, *LLaMA 3.2* [13]) to refine the segmentation, ensuring blocks are informative and coherent. Images are extracted using the *PyMuPDF* Python package [14]. Both text blocks and images are embedded into a shared semantic space using CLIP [15]. These embeddings, together with metadata (including company name, document name, page number, and block/image ID), are stored in a vector database to enable efficient semantic retrieval.

Given a sustainability objective query, the system embeds the query using CLIP and retrieves the most similar text blocks and images from the company's vector database using cosine similarity. For each retrieved text block b_i, the system also retrieves neighboring blocks within a window of size W, i.e., $B_{\text{previous}} = [b_{i-W}, b_{i-W+1}, \ldots, b_{i-1}]$ and $B_{\text{next}} = [b_{i+1}, b_{i+2}, \ldots, b_{i+W}]$, to preserve contextual coherence. We use an open-source LLM (in the current prototype, *LLaMA 3.2*) to generate a final verification report using all the retrieved textual evidence. The model is instructed to produce a concise, factual statement that synthesizes the multimodal evidence in favor of or against the queried sustainability objective. Relevant images are displayed alongside the generated text in the user dashboard.

III. Evaluation

Our evaluation experiments aim to compare the effectiveness of our fact-checking system against a baseline on a sample of our dataset.

Dataset. We chose 5 companies from the technology, healthcare, energy, food, and retail domains. For each company, we collected the last 5 annual sustainability reports from 2020–2024. We selected a sample of 50 extracted sustainability objectives from these reports for the fact-checking task. We use this diverse data sample to test the robustness of fact-checking systems across various domains.

TABLE I
SYSTEM EFFECTIVENESS IN COMPARISON TO THE BASELINE.

Measure	Baseline	Our System
ROUGE-1	0.06	**0.21**
ROUGE-2	0.01	**0.16**
ROUGE-L	0.05	**0.19**
BERTScore	0.80	**0.85**
Pairwise Human Preference Rate	0.19	**0.81**

Baseline. Given a sustainability objective query, the baseline approach treats the text of each page as a text block. It uses the same LLM (i.e., *LLaMA 3.2* [13]) and prompt as our system to produce a multimodal verification report. It also includes a sorted list of images based on the lexical similarity of their surrounding text to the given sustainability objective. This baseline serves to evaluate the effectiveness of our system design, as it excludes our multimodal RAG component.

Evaluation measures. We evaluate using syntactic, semantic, and human metrics. *ROUGE-N* [16] measures n-gram overlap between the generated verification report and the reference sustainability objective to assess their syntactic similarity. *BERTScore* [17] computes the cosine similarity of embeddings to assess their semantic similarity. In *Pairwise Human Preference Rate* [18], two independent evaluators compared verification reports from our system and the baseline in a blind, side-by-side experiment to determine which better supports fact-checking of the given sustainability objective. We report the proportion of times each approach was preferred to indicate effectiveness in real-world human evaluation. For each evaluation measure, we report the mean across all samples and evaluators. For the sake of readability, we omit the standard errors as they are always small numbers close to zero ($< 1\%$).

Results. We compare the effectiveness of our system with the baseline in Table I. Our system outperforms the baseline significantly and consistently across all evaluation metrics, reflecting our careful system design. The higher *ROUGE-N* score indicates that our RAG component retrieves evidence effectively for the given sustainability objective. The higher *BERTScore* shows that our generated verification report is also semantically relevant to the objective. Finally, the superior *Pairwise Human Preference Rate* demonstrates that our system better supports users in real-world fact-checking scenarios, thanks to its multimodal RAG component.

IV. DEMONSTRATION

We demonstrate our sustainability objective fact-checking system to illustrate how it supports users across various fact-checking scenarios. The audience for our system includes a *wide range of users*, such as *sustainability domain experts* seeking to audit companies, *investors* looking to evaluate company values, *vigilant consumers* who want to choose environmentally friendly products and services, and *technical enthusiasts* interested in innovative applications of emerging technologies.

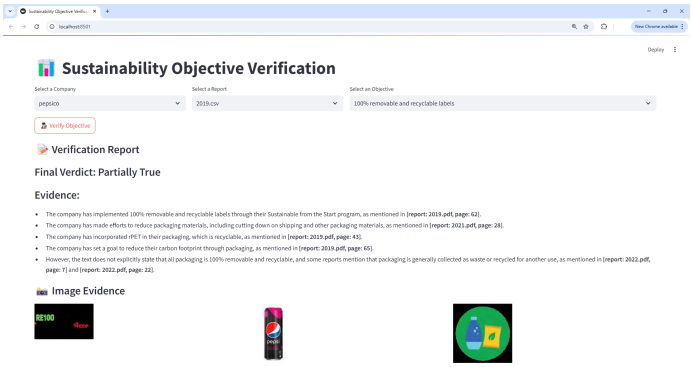

Fig. 2. The sustainability objective fact-checking dashboard.

Fact-checking sustainability objectives. The most common scenario involves a user seeking to fact-check the sustainability claims made by a specific company. Figure 2 shows our fact-checking dashboard designed to support the user in this process. The user first chooses a company and a sustainability report for that company from internal records. Next, our system uses GoalSpotter [5] to extract and list the sustainability objectives mentioned in the selected report. The user can select any of these objectives and click the "Verify Objective" button. The system then processes the company's historical reports, retrieves multimodal evidence, and generates a verification report for the user, including a final verdict and a list of supporting/contradicting evidence. As shown in the figure, for each retrieved piece of evidence, the system also provides reference metadata so the user can locate the original report and page number for further investigation. The system also retrieves and displays relevant images.

Comparison with the baseline. The next scenario compares the performance of our sustainability objective fact-checking system with the baseline presented earlier. We compare the output of our system side by side with the baseline to demonstrate how our system achieves higher effectiveness and mitigates common LLM limitations, such as hallucinations, through its multimodal RAG component.

Company comparison. Another scenario is when the user wants to compare two companies based on a specific sustainability objective, such as "achieving a net-zero carbon footprint." This is especially valuable for consumers who have multiple options for products or services and wish to choose the more sustainable provider. To do this, the user should extract and validate all relevant evidence for the given objective from both companies and compare the results side by side. We present our company comparison dashboard, which enables users to select two companies and a sustainability objective to compare their performance side by side.

V. RELATED WORK

Our system retrieves multimodal evidence using a RAG component to verify sustainability objectives. Therefore, we first review sustainability objective detection approaches. Next, we discuss fact-checking methods to verify sustainability objectives. Finally, we look at RAGs.

Sustainability objective detection. Traditionally, domain experts have manually detected sustainability objectives as the first part of greenwashing detection [6]. This process involves manual analysis of many sustainability reports of a company over the years and cross-checking them against standard frameworks and criteria [8], [1], [4]. Recent developments have shown the possibility of automating this task effectively [19]. These approaches usually segment sustainability reports into smaller text units to classify them into objective/non-objective categories using recent transformer models [5], [20], [21].

Sustainability objective detection is an upstream task for our fact-checking problem. The input sustainability objectives to our fact-checking system can be extracted using either an automated or manual approach. In our current prototype, we use GoalSpotter [5] to automatically extract sustainability objectives, since our main problem is to support domain experts with fact-checking these input objectives.

Fact-checking systems. Once a claim is spotted, a fact-checking process needs to evaluate it. In journalism, the goal of fact-checking is to verify the veracity of a claim using its textual information [22]. Claim verification consists of several steps, including relevant document retrieval, evidence extraction, stance detection, verdict prediction, and justification production [23], [24], [25]. The current dominant approach for many of these tasks is to fine-tune large pre-trained language models [24]. Due to the complexity of the process, a human-in-the-loop is a recommended setting for fact-checking to assess the final output of the fact-checking system [26]. Examples of fact-checking systems in political science include FactCheckBureau [27] and RAGAR [28]. Recent public benchmark datasets, such as CheckThat![29] and NEWSCLAIMS[30], support researchers in developing political fact-checking systems.

However, existing political fact-checking systems and benchmark datasets are not directly applicable, as the sustainability data domain and claims differ significantly from those in politics. Existing sustainability fact-checking research either only analyzes sustainability claims and verification practices [31], [32] or focuses on narrow scopes, such as classifying climate-related statements by their level of truthfulness [33]. That is why we have developed a new, dedicated fact-checking system for verifying sustainability objectives. To our knowledge, our system is the first approach to automatically retrieve multimodal evidence to support domain experts in fact-checking of all kinds of sustainability objectives.

Retrieval-Augmented Generation (RAG). RAG is an approach that combines retrieved content with generative models to produce grounded and context-based outputs [34]. In this framework, a retriever fetches relevant texts/images based on a given query, and a generator, typically a transformer-based large language model, produces a response conditioned on both the query and the retrieved evidence. RAG has been used in various problems, such as question answering [35], scientific claim verification [36], and political fact-checking [28].

We adapt the RAG paradigm to the sustainability domain to generate verification reports for specified sustainability objectives using retrieved multimodal evidence.

VI. Conclusion

We have presented a novel sustainability objective fact-checking system that retrieves multimodal evidence to generate verification reports. Our experiments demonstrate that the system effectively verifies sustainability objectives based on syntactic, semantic, and human evaluation metrics. As a future research direction, we plan to extract finer-grained details of sustainability objectives to further enhance the fact-checking process.

References

[1] S. V. de Freitas Netto, M. F. F. Sobral, A. R. B. Ribeiro, and G. R. d. L. Soares, "Concepts and forms of greenwashing: A systematic review," *Environmental Sciences Europe*, vol. 32, no. 1, pp. 1–12, 2020.

[2] United Nations, "Sustainable development goals," https://sdgs.un.org/goals, 2025, accessed: 25.09.2025.

[3] United Nations Framework Convention on Climate Change (UNFCCC), "The paris agreement," https://unfccc.int/process-and-meetings/the-paris-agreement, 2015, accessed: 25.09.2025.

[4] W. Moodaley and A. Telukdarie, "Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review," *Sustainability*, vol. 15, no. 2, p. 1481, 2023.

[5] M. Mahdavi, R. Baghaei Mehr, and T. Debus, "Combat greenwashing with goalspotter: Automatic sustainability objective detection in heterogeneous reports," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 4752–4759.

[6] TerraChoice, "The sins of greenwashing: Home and family edition," *Underwriters Laboratories*, 2010.

[7] Greenwashing Index, "About greenwashing," https://web.archive.org/web/20130426232159/http://www.greenwashingindex.com/about-greenwashing, 2025, accessed: 25.09.2025.

[8] L. Gatti, P. Seele, and L. Rademacher, "Grey zone in–greenwash out. a review of greenwashing research and implications for the voluntary-mandatory transition of csr," *International Journal of Corporate Social Responsibility*, vol. 4, no. 1, pp. 1–15, 2019.

[9] M. Gräuler and F. Teuteberg, "Greenwashing in online marketing–investigating trust-building factors influencing greenwashing detection," *Multikonferenz Wirtschaftsinformatik*, pp. 1359–1366, 2014.

[10] N. Nemes, S. J. Scanlan, P. Smith, T. Smith, M. Aronczyk, S. Hill, S. L. Lewis, A. W. Montgomery, F. N. Tubiello, and D. Stabinsky, "An integrated framework to assess greenwashing," *Sustainability*, vol. 14, no. 8, p. 4431, 2022.

[11] N. R. Sahoo, A. Saxena, K. Maharaj, A. A. Ahmad, A. Mishra, and P. Bhattacharyya, "Addressing bias and hallucination in large language models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024, pp. 73–79.

[12] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," *arXiv preprint arXiv:2509.04664*, 2025.

[13] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[14] Artifex Software Inc., "Pymupdf: High-performance python library for data extraction, analysis, conversion and manipulation of pdf (and other) documents," 2025, accessed: 25.09.2025. [Online]. Available: https://pypi.org/project/PyMuPDF

[15] A. Radford, J. W. Kim, M. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021.

[16] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[18] Y. Liu, H. Zhou, Z. Guo, E. Shareghi, I. Vulić, A. Korhonen, and N. Collier, "Aligning with human judgement: The role of pairwise preference in large language model evaluators," *arXiv preprint arXiv:2403.16950*, 2024.

[19] T. Cojoianu, A. G. Hoepner, G. Ifrim, and Y. Lin, "Greenwatch-shing: Using ai to detect greenwashing," *AccountancyPlus-CPA Ireland*, 2020.

[20] V. Woloszyn, J. Kobti, and V. Schmitt, "Towards automatic green claim detection," in *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, 2021, pp. 28–34.

[21] D. Stammbach, N. Webersinke, J. A. Bingler, M. Kraus, and M. Leippold, "A dataset for detecting real-world environmental claims," *Center for Law & Economics Working Paper Series*, vol. 2022, no. 07, 2022.

[22] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "Fully automated fact checking using external sources," *arXiv preprint arXiv:1710.00341*, 2017.

[23] A. Hanselowski, C. Stab, C. Schulz, Z. Li, and I. Gurevych, "A richly annotated corpus for different tasks in automated fact-checking," *arXiv preprint arXiv:1911.01214*, 2019.

[24] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, p. e12438, 2021.

[25] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

[26] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino, "Automated fact-checking for assisting human fact-checkers," *arXiv preprint arXiv:2103.07769*, 2021.

[27] O. Balalau, P. Bertaud-Velten, Y. El Fraihi, G. Gaur, O. Goga, S. Guimaraes, I. Manolescu, and B. Saadi, "Factcheckbureau: Build your own fact-check analysis pipeline," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 5185–5189.

[28] M. A. Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletić, "Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models," *arXiv preprint arXiv:2404.12065*, 2024.

[29] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino *et al.*, "Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media," *Conference and Labs of the Evaluation Forum (CLEF) Working Notes*, vol. 2696, 2020.

[30] R. G. Reddy, S. Chetan, Z. Wang, Y. R. Fung, K. Conger, A. Elsayed, M. Palmer, P. Nakov, E. Hovy, K. Small *et al.*, "Newsclaims: A new benchmark for claim detection from news with attribute knowledge," *arXiv preprint arXiv:2112.08544*, 2021.

[31] H. Kang and J. Kim, "Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods," *Applied Sciences*, vol. 12, no. 11, p. 5614, 2022.

[32] H. T. Vu, A. Baines, and N. Nguyen, "Fact-checking climate change: An analysis of claims and verification practices by fact-checkers in four countries," *Journalism & Mass Communication Quarterly*, vol. 100, no. 2, pp. 286–307, 2023.

[33] M. Leippold, S. A. Vaghefi, D. Stammbach, V. Muccione, J. Bingler, J. Ni, C. C. Senni, T. Wekhof, T. Schimanski, G. Gostlow *et al.*, "Automated fact-checking of climate claims with large language models," *npj Climate Action*, vol. 4, no. 1, p. 17, 2025.

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[35] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781.

[36] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: Verifying scientific claims," *arXiv preprint arXiv:2004.14974*, 2020.