SNOOP Method: Faithfulness of Text Summarizations for Single Nucleotide Polymorphisms

Wolfgang Maass,[1] **Cicy K. Agnes,** [1] **Maxx R. Rahman** [1] **Jonas S. Almeida** [2]

[1] German Research Center for Artificial Intelligence (DFKI), 66123 Saarbruecken, Germany
[2] Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, Rockville, MD 20850, USA.
wolfgang.maass@dfki.de, cicy.agnes@dfki.de, amin.harig@dfki.de, jonas.dealmeida@nih.gov

Abstract

Time pressures and a heavy workload often limit a physician's ability to keep up with the increasing number of scientific publications. It is hoped that text summarization by large language models (LLM) can help practitioners quickly identify essential publications. However, it is unknown whether LLMs have been trained on scientific publications in medicine and whether summaries are faithful or even caused by hallucinations. We present the SNOOP method, which uses transformer model embeddings to assess fidelity and hallucinations for different types of LLM summaries and provides an integrated view of results that can be quickly assessed by physicians. In the context of genomic medicine, we present results on the performance of SNOOP-enhanced LLMs.

Introduction

According to PubMed, around 50.000 papers have been published in 2020 on Covid-19 (Else 2020). In general, about 1Mio. papers are included in PubMed every year, i.e. two papers per minute (Landhuis 2016). Automatic summarization of medical publications has been a research topic for decades (Fan et al. 2006) but only with the advent of large-language models (LLM) it has reached the interest of millions, including medical researchers and practitioners. Recent studies show the high quality of LLMs-generated summarizations that are even comparable with humans (Zhang et al. 2023). LLM promises a professional-grade means for the identification of relevant results in older but also recent publications. However, research on the quality of LLM-based text summarizations reduced euphoria by eliciting frequent hallucinations (Maynez et al. 2020). Furthermore, LLM service providers constantly redesign capabilities by restricting allowed requests (aka prompts) but also restricting responses while it is unknown which scientific publications were used for training. This unsecured state means that medical researchers and physicians cannot trust LLM-based summarizations without an assessment of the quality of text summarizations.

Faithfulness and factuality are proposed as key indicators for the quality of LLM-based text summarizations (Maynez et al. 2020). Faithfulness is defined as staying consistent and truthful to the provided source — an antonym to "hallucination" (Ji et al. 2023). In this paper, we understand faithfulness as the characteristic of a summarization that its information is directly grounded in the input text. For nouns, this means that they are directly found in the input text on syntactial level. For instance, the term '*colon cancer*' in a text summarization is being used in the input text. For medical publications, *direct faithfulness* is important because medical terms shall remove ambiguities and increase precision. Factuality refers to the quality of being actual or based on fact (ibid.). This extends faithfulness by allowing more general concepts and generalizing descriptions beyond the actual content of a publication. Some researchers argue for removing factuality (Dong et al. 2020).

In this paper, we will focus on direct faithfulness of text summarizations and their use in the domain of genomics medicine. Genomics medicine investigates the combination of genetic and environmental factors causing complex diseases such as heart disease, asthma, diabetes, and cancer (Feero, Guttmacher, and Collins 2010). Integrating genomic, lifestyle, ancestry, and other sources of information is a complex tasks that exceeds cognitive resources of physicians in daily practice. Today, physicians are overwhelmed by the increasing number of publications that present results of genome-wide association studies (GWAS), which scan the entire genome for variations that are associated with a particular disease or trait (Fatumo et al. 2022). Therefore, physicians require trustworthy tools that summarize and extract relevant information that can be used for personalized medicine.

In this paper, we present an approach for assessing and increasing the faithfulness of text summarizations in the medical domain. Our approach integrates LLM for text summarizations, text embedding models, and unsupervised models for clustering. We use these AI technologies for comparing LLM-generated summarizations with abstracts in original papers and use out-of-domain documents for quality assessment. The approach will be applied to polygenic risk scores and associated single nucleotide polymorphisms (SNPs). Our research is driven by the following questions:

- **RQ1:** How to identify hallucinations in LLM summarizations of medical publications?
- **RQ2:** How to increase faithfulness in summarizations of medical publications?

State of the Art

Polygenic risk scores

Recently, polygenic risk scores (PRS) have become an important tool in preventive medicine for the interpretation of patient risks (Chatterjee, Shi, and García-Closas 2016). PRS uses statistical methods for estimating a person's genetic risk for a particular disease or trait based on multiple genetic variants which is feasible on a large scale due to commercial genome sequencing (Krier, Kalia, and Green 2022). These variants are identified through analysis of data obtained by Genome-wide association studies (GWAS). PRS are based on linear combinations of weighted scores associated with Single Nucleotide Polymorphism (SNP). An SNP is a type of genetic variation that occurs when a single nucleotide at a specific position in the genome differs between individuals in a reference population (Shastry 2002).

Experts analyze results, distill key data, and store it in web-based databases such as dbSNP[1], ClinVar [2], dbVar[3], Atlas [4] and many other databases.

The dbSNP database currently contains 900 million records on genetic variations. The NHGRI-EBI Genome-wide association studies (GWAS) Catalog contains metadata for $> 45,000$ published GWAS across $> 5,000$ human traits and $> 40,000$ full p-value summary statistics datasets. It is evident that automatic text summarization is of utmost importance for genomic researchers but also medical professionals (Landhuis 2016).

Quality of text summarizations

Extractive summarization selects salient words and phrases while abstractive summarization uses embeddings and transformations for generating compressed paraphrases of input documents (Radford et al. 2019). Automatic text summarization is the process of balancing the trade-off between reducing a document while preserving essential information content and meaning (Pilault et al. 2020). The quality of text summarization by LLMs is assessed by the comparison with reference texts measured by scores, such as BLEU and ROUGE (Durmus, He, and Diab 2020; Lin 2004). Faithfulness measures the amount of information in a summarization that is supported by input documents (Maynez et al. 2020). *Direct faithfulness* is given by the syntactic presence of information in input texts while *abstract faithfulness* assesses the semantic support of summarization information by input texts. Large language models (LLM), such as chatGPT (Qin et al. 2023), have proven to generate abstractive summarization with high levels of fluency and coherence (Pilault et al. 2020) while occasionally exhibiting flawed results that lack faithfulness and factuality (Maynez et al. 2020). Maynez et al. found that more than 70% of all single-sentence abstractive summarizations suffer from hallucinations that add information not present in the input text (Maynez et al. 2020). Analog to faithfulness, two types of hallucinations are distinguished. *Direct hallucination* is information that is not present in input texts on item-level by assuming a closed-world assumption. For instance, creating a reference to a non-existing publication. *Abstract hallucination* is information that is not within the conceptual scope of input texts. For instance, giving information on interactions between SNPs that do not exist. In the following, we will focus on direct hallucinations and direct faithfulness.

Embeddings

Embedding techniques such as Word2Vec (Mikolov et al. 2013) capture semantic relationships between words. Text preprocessing techniques including stop word removal and lemmatization improve the quality of text embeddings. Similarly, the quality of embeddings is improved by using abstracts instead of keywords (Alexandrov, Gelbukh, and Rosso 2005).

In summary, LLM-based summarizations of publications and general topics, such as SNPs, suffer from major quality issues, such as lack of faithfulness and hallucinations. In the following, we present the *SNOOP method* that allows quality assessment of LLM-based summarizations of documents ("*Give a summary of publication Grampp et al 2017 Multiple renal cancer susceptibility polymorphisms modulate the HIF pathway*") but also summarizations of supporting documents on abstract topics, called references ("*Give a summary of publications for rs7105934*").

SNOOP Method

The SNOOP method improves direct faithfulness of summarizations and identifies direct hallucinations. Summarization is performed on two levels: (1) references, and (2) documents. References are markers for topics, such as the name of a person in Wikipedia. Documents are linked to references if there is a reliable source that allows this linking. For example, a list of documents linked to a reference in Wikipedia or an SNP name.

Since in many cases it is not known what data was used to train an LLM, it is unclear whether the documents of interest or any information about the references of interest were used for training. The SNOOP method generates LLM summarizations for references and associated documents and assesses potential hallucinations and faithfulness. Faithfulness is assessed relative to documents that are *qualified knowledge*, i.e., high-level scientific publications, such as publications of GWAS studies on SNPs in high-end journals. We assume that documents of qualified knowledge are faithful descriptions of references. Documents of qualified knowledge are used for assessing the faithfulness of LLM summarizations. Because understanding scientific documents containing qualified knowledge is often not suitable for use in the daily practice of physicians, LLM summarizations provide a means of verifying that faithfulness is maintained at a satisfactory level. For this, associated documents from sources of qualified knowledge are selected and used as prompts for LLM summarization. References are conceptual entities, e.g., person name, drug name, or SNP name. LLM summaries of a set of related publications define a clear boundary, while LLM summaries of references refer to undefined

[1]https://www.ncbi.nlm.nih.gov/snp/)

[2]https://www.ncbi.nlm.nih.gov/clinvar/

[3]https://www.ncbi.nlm.nih.gov/dbvar/

[4]https://atlas.ctglab.nl/

Algorithm 1: SNOOP method

Input: *reference* (e.g., SNP), *associations* (e.g., dbSNP), *out-of-domain-docs*

Output: prompt reply

1: *refSum* = summarization(*reference*)
2: *abstracts* = SelectAbstracts(*publicationsByAssociation*)
3: *pubsSums* = GenSum(*publicationsByAssociation*)
4: *summary* = summarization(*abstracts*)
5: *texts* = *pubsSums* ∪ *refSum* ∪ *summary* ∪ *out-of-domain-docs*
{*Embedding and Clustering*}
6: *emb* = Embeddings(texts)
7: *dist* = distance(*emb*)
8: *c* = Clustering(*emb*)
9: Visualize(*c*)
{*Halluzination Assessment*}
10: *support* = GenRefSupport(refSum, reference)
11: *h* = hallucination(support)
{*Faithfulness assessment: distance*}
12: *f* = faith($c_{abstracts}$, $c_{refSum} \cup c_{pubsSums}$, *support*, *h*)
13: *reply* = CreateReply(*h, f*)
14: **return** *reply*

content whose range is drawn internally only through statistical embeddings and similarities by the LLM. Therefore, it is important to assess whether an LLM has the ability to abstract knowledge from documents to conceptual entities. The output of the SNOOP method provides physicians with three types of summarizations including indicators for faithfulness and hallucinations: (1) integrated summarization over all documents of qualified knowledge, (2) summarization of each document of qualified knowledge, and (3) summarization of knowledge about the reference. A summarization is omitted if it contains direct hallucinations or content with insufficient direct faithfulness.

Step 1 of the SNOOP method generates an LLM summarization for the reference (cf. Algorithm 1). Abstracts of publications associated with the reference are extracted (step 2), and corresponding summarizations and summarization of all abstracts of associated documents are generated by an LLM (GenSum in step 3). All texts are embedded, distances are measured (e.g., cosine distance), clusters are determined (e.g., UMAP and dbSCAN), and clusters are visualized (steps 5-9). Step 10 identifies direct hallucinations by (1) testing for the existence of references given by LLM (*support*) and (2) by testing for the occurrence of references in documents (Steps 10 - 11). If one of these criteria fails, a reference is identified as a direct hallucination. For all text items that are not hallucinated, the function *faith* determines the faithfulness between qualified texts ($c_{abstracts}$) and embeddings of the reference summarization (c_{refSum}) and document summarizations (*pubsSums*). All texts that fall within the range of distances of cluster $c_{abstracts}$ and that are non-hallucinatory will be integrated into *reply* (step 13).

The SNOOP method is agnostic to LLMs, such as chatGPT or Bard, and embedding methods, such as SBERT (Reimers and Gurevych 2019) and USE (Cer et al. 2018). It can be used for different domains as long as documents of qualified knowledge exist.

Study

The SNOOP method will be illustrated by a study that starts with a European patient whose genome shows a $G > A$ variation at SNP rs7105934. About 8% of all Europeans carry this variation while GWAS studies show that this variation has associations with renal cancer. Therefore the physician wants to update her view of SNP rs7105934 by using summarizations provided by a SNOOP-enhanced LLM (here: chatGPT). In this section, we describe the implementation of the proposed method for hallucination and faithfulness assessment and its application to SNP rs7105934.

Datasets

The SNP number is used as an index for retrieving all the registered publications from dbSNP. We retrieved all the publications associated with SNP rs7105934 and some publications from different domains.

- 10 publications associated with rs7105934 in dbSNP, e.g. (Grampp et al. 2017): *associated*
- 1 summary for a movie (Pulp Fiction): *out-of-domain*
- 15 publications on philosophy: *out-of-domain*
- 13 publications associated with artificial intelligence in arXiv: *out-of-domain*

Embedding Models

- **SBERT:** (Reimers and Gurevych 2019) Sentence-BERT is a text embedding model that enhances the representation of sentences by capturing their semantic information. It utilizes a triplet network architecture to fine-tune the popular BERT model for sentence-level tasks and produce dense embeddings.
- **USE:** (Cer et al. 2018) Universal Sentence Encoder is a text embedding model that converts sentences into high-dimensional vectors and captures their semantic information. It is trained on a large corpus of text and can generate embeddings for a wide range of languages and sentence lengths.
- **Open-AI:** It is trained on a vast amount of text data and utilizes a combination of deep learning techniques and transformer-based architectures to generate powerful embeddings. These embeddings capture the semantic information and contextual understanding of sentences.

Performance Measure

- **Cosine distance:** (Huang 2008) The cosine similarity is used as the evaluation metrics for faithfulness assessment. It is a measure of similarity defined in an inner product space.

$$cos\theta = \frac{x \cdot y}{|x||y|}$$

where x and y represent the embedding vectors of the two texts.

Hallucination Test

We applied the proposed SNOOP method to find the hallucination in LLM summarizations for the reference, each publication, and the set of publications. Reference SNP rs7105934 was given as input to LLM and asked "*Give the key publications for rs7105934?*". LLM returned a list of 5 publications associated with the given reference. Additionally, summarizations were generated for each of the 10 publications[5] that are registered in dbSNP for the reference SNP rs7105934 (qualified knowledge). Additionally, a summarization is generated for all publications in summary. After embedding and clustering (cf. Figure 1), all summarizations are tested for direct hallucinations.

Summarizations for publications are non-hallucinated by definition because they are qualified knowledge but the supporting publications for the reference summarization are not. According to step 10 and 11, SNOOP checks for the existence of the publication given by LLM on Google Scholar[6], and we finds that two of them do not even exist. It shall be noted that the doi of these two publications directs to publications in Pubmed[7] and JACC[8] databases with different titles and authors, both without connections to the reference rs7105934. Here is a list of these two publications completely hallucinated by LLM: [9]

- Lü M, Yang J, Chen Y, et al. 2011. Genetic variants on chromosome 4q25 are associated with a risk of atrial fibrillation: evidence from two prospective cohort studies. J Am Coll Cardiol, 58(7):696-704. doi:10.1016/j.jacc.2011.03.042
- Cicek MS, Liu J, Casey G, et al. 2015. Colorectal cancer risk variants at 8q23.3 and 11q23.1 are associated with disease phenotype in African Americans. Carcinogenesis, 36(5):573-579. doi:10.1093/carcin/bgv029

The next step checks for the occurrence of reference rs7105934 in the other three publications given by LLM. None of these articles mentions *rs7105934* either in their title, abstract, or body. Therefore, all supporting documents for the reference summarization are classified as being directly hallucinated so the summarization is also classified as being unfaithful. This also means that chatGPT in its current version has problems with finding proper conceptual abstractions, i.e., building a relationship between instances given by documents and concepts given by references.

Faithfulness Assessment

The following texts are used for faithfulness assessment:

1. Abstracts of documents from qualified knowledge (10)
2. Summarization of all documents from qualified knowledge (1)
3. Summarization of out-of-domain documents
4. Summarization of the reference rs7105934 generated by chatGPT
5. Summarizations of the three existing but hallucinated documents

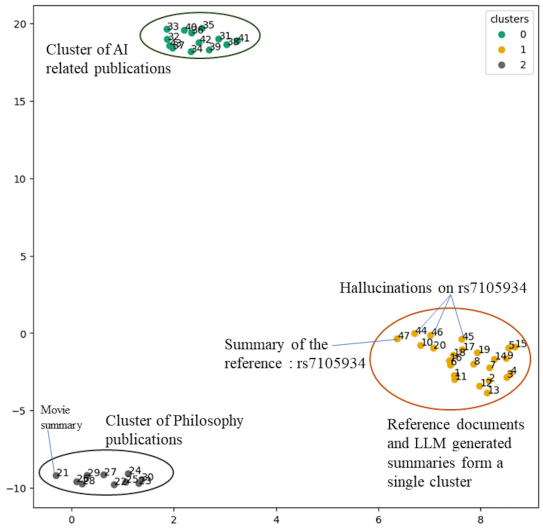

Figure 1: Clustering analysis of all the abstracts and LLM summaries. Point 1 is LLM summary of (Grampp et al. 2017), 2-11 denotes abstracts of associated publications, 12-40 denotes abstracts of out-of-domain publications and 41-52 represents LLM summary (including hallucination on reference rs7105934) on associated publications in dbSNP.

All texts were embedded by using three different embedding models: SBERT (vector: length 768), USE(vector length 512), and Open-AI (vector: length 1536). To project these vectors into a two-dimensional space, we used a manifold learning technique, i.e., UMAP (McInnes et al. 2018) which performs non-linear dimension reduction and reduces these vectors into a vector of length 2. Next, we applied the DBSCAN algorithm (Ester et al. 1996) on these reduced vectors to perform cluster analysis. It is a density-based spatial clustering which works well since we have different density regions corresponding to associated and out-of-domain publications. In addition, we expect an abstract and LLM summary of the same publication to be a part of the same cluster in order to be a faithful summary, and the visualization of these clusters in two-dimensional space helps for the direct assessment. To quantify faithfulness, we measure the cosine distance between the textual embeddings of the abstract and the LLM summary of 10 associated publications.

Fig 1 shows the association of all the publications to give a direct faithfulness assessment on the LLM summaries. We observed the formation of three clusters where each publication is associated with one of them. In the first cluster, we observed that the LLM summaries and the SNP abstracts of the associated publications are clustered together, which gives a faithfulness check and tells us about their similarity. This cluster consists of publications related to SNP rs7105934,

[5]https://www.ncbi.nlm.nih.gov/snp/rs7105934#publications
[6]https://scholar.google.com/
[7]https://pubmed.ncbi.nlm.nih.gov/
[8]https://www.jacc.org/
[9]Note: ChatGPT was recently updated on 12 May 2023 and now blocks any requests for medical publications. However, the GPT model has not changed.

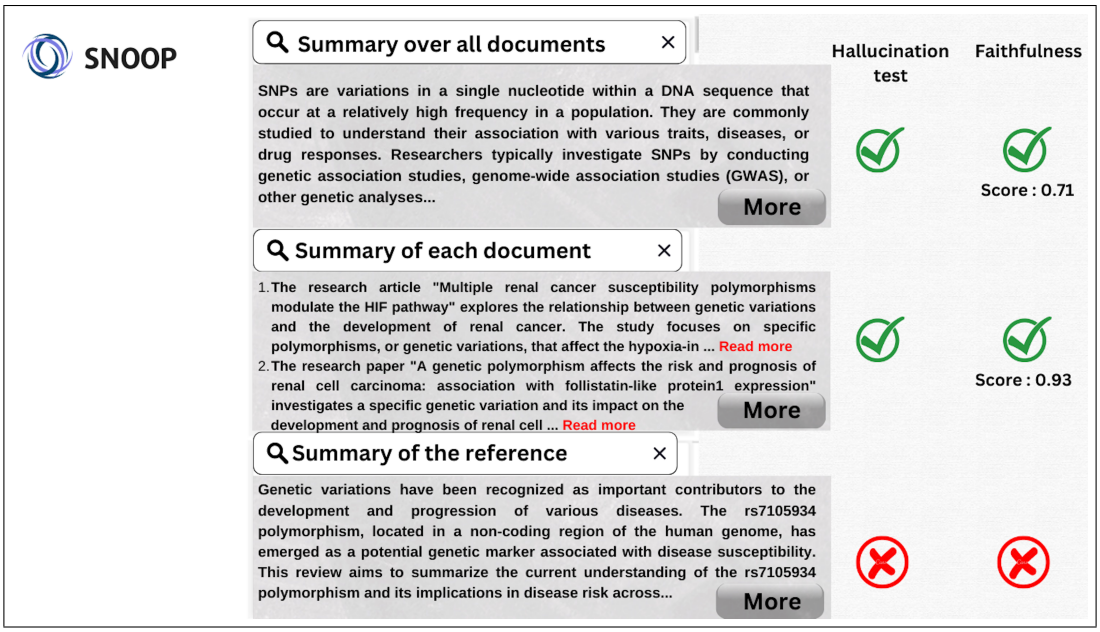

Figure 2: SNOOP System: summary of results for reference rs7105934.

i.e., 10 abstracts and 10 LLM summaries of associated publications. It is interesting to note that the 3 hallucinated articles are also part of this same cluster which tells that they belong to the domain of genomic medicine. On the other hand, we found that the out-of-domain publications are clustered separately into philosophy and AI-related publications, whereas the movie summary is clustered into philosophy.

For the quantified measure of direct faithfulness on the associated publications, we have Table 1 showing the cosine distance between the SNP abstracts and LLM summarizations. These distances are calculated for each embedding model separately to show the impact of the textual embeddings. When using the Open-AI embedding model, the embeddings of the abstracts and summaries related to SNP rs7105934 are significantly closer than the sentence-BERT and universal sentence encoder because the LLM model (chatGPT) uses the same textual embeddings to convert the text and generate summaries. Therefore, we observe higher cosine values (> 0.9) between the embeddings of SNP abstracts and LLM summarization. The similarity between the textual embeddings of the LLM summary of (Grampp et al. 2017) and the SNP abstract is 0.876 for the sentence-BERT model, which indicates a high overlap between the LLM-generated summary and the authored abstract of the publication.

Fig. 2 summarises the results of the SNOOP method for the hallucination and faithfulness assessment of LLM summarization. First, we have the LLM summary over all the associated publications together, which passes the hallucination test since rs7105934 is mentioned in their summary. For the faithfulness assessment, we calculated the cosine distance between the point representing the LLM summary and the centroid over all the points representing the associated publications in the UMAP space and achieved a score of 0.71. Next, we have the hallucination test and the faithfulness assessment for the LLM summary of each of the associated documents related to rs7105934. Finally, we present the summarization for the reference rs7105934 which was identified as being hallucinated due to hallucinated supporting documents (5 out of 5). The SNOOP interface provides a qualified overview of faithful summarizations for a given SNP reference. It provides a generated overview on all documents of qualified knowledge, summarizations of all documents, and a summary of the reference plus a qualification of hallucination scores and faithfulness scores.

Pub	SBERT	USE	Open-AI
1	0.876	0.676	0.924
2	0.889	0.690	0.938
3	0.892	0.605	0.932
4	0.807	0.539	0.910
5	0.922	0.626	0.938
6	0.711	0.707	0.923
7	0.786	0.627	0.936
8	0.843	0.595	0.933
9	0.830	0.641	0.946
10	0.872	0.531	0.919

Table 1: Cosine distance between abstracts and LLM summaries of associated publications when different embedding models are used.

Conclusion

Using the SNOOP method, we presented a method for identifying direct hallucinations and evaluating the direct faith-

fulness of LLM summarization. SNOOP uses documents of qualified knowledge as a reference against which LLM summarizations are evaluated. As illustrated by the study, faithfulness tests based on embeddings are insufficient for the identification of hallucinations. Hallucinations are identified by background checks on the existence of documents and the occurrence of reference markers in documents. Our study shows promising performance results of the SNOOP method in the context of genomic medicine that will be extended in the future. We will extend the SNOOP method to abstract hallucinations and abstract faithfulness in collaboration with medical experts.

References

Alexandrov, M.; Gelbukh, A.; and Rosso, P. 2005. An approach to clustering abstracts. In *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005. Proceedings 10*, 275–285. Springer.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Strope, B.; and Kurzweil, R. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174. Brussels, Belgium: Association for Computational Linguistics.

Chatterjee, N.; Shi, J.; and García-Closas, M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7): 392–406.

Dong, Y.; Wang, S.; Gan, Z.; Cheng, Y.; Cheung, J. C. K.; and Liu, J. 2020. Multi-fact correction in abstractive text summarization. *arXiv preprint arXiv:2010.02443*.

Durmus, E.; He, H.; and Diab, M. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Else, H. 2020. COVID IN PAPERS. *Nature*, 588(24/31): 553.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Fan, W.; Wallace, L.; Rich, S.; and Zhang, Z. 2006. Tapping the power of text mining. *Communications of the ACM*, 49(9): 76–82.

Fatumo, S.; Chikowore, T.; Choudhury, A.; Ayub, M.; Martin, A. R.; and Kuchenbaecker, K. 2022. A roadmap to increase diversity in genomic studies. *Nature medicine*, 28(2): 243–250.

Feero, W. G.; Guttmacher, A. E.; and Collins, F. S. 2010. Genomic medicine—an updated primer. *New England Journal of Medicine*, 362(21): 2001–2011.

Grampp, S.; Schmid, V.; Salama, R.; Lauer, V.; Kranz, F.; Platt, J.; Smythies, J.; Choudhry, H.; Goppelt-Struebe, M.; Ratcliffe, P.; Mole, D.; and Schödel, J. 2017. Multiple renal cancer susceptibility polymorphisms modulate the HIF pathway. *PLoS genetics*, 13: e1006872.

Huang, A.-L. 2008. Similarity Measures for Text Document Clustering.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Krier, J. B.; Kalia, S. S.; and Green, R. C. 2022. Genomic sequencing in clinical practice: applications, challenges, and opportunities. *Dialogues in clinical neuroscience*.

Landhuis, E. 2016. Scientific literature: Information overload. *Nature*, 535(7612): 457–458.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Pilault, J.; Li, R.; Subramanian, S.; and Pal, C. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9308–9319.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shastry, B. S. 2002. SNP alleles in human disease and evolution. *Journal of human genetics*, 47(11): 561–566.

Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; and Hashimoto, T. B. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.