Advanced RAG-LLM Prototype AI on PubMed for Cardiac Health

Luuk P.A. Simons, Pradeep K. Murukannaiah, Budi S. Han, Mark A. Neerincx

Delft University of Technology, Faculty of EEMCS, Delft, the Netherlands
l.p.a.simons@tudelft.nl, P.K.Murukannaiah@tudelft.nl, b.s.han@student.tudelft.nl, M.A.Neerincx@tudelft.nl

DOI
https://doi.org/10.18690/um.fov.4.2025.17

ISBN
978-961-286-998-4

Healthy lifestyle behaviours are effective in preventing and treating cardiovascular disease. However, the growing body of scientific literature and the prevalence of conflicting studies make it challenging for healthcare practitioners and patients to stay informed. Large Language Models (LLMs), combined with Retrieval-Augmented Generation (RAG), enable automated claim verification and summarization. We enhanced RAG-LLM with extra modules and evaluated performance. Inclusion-Criteria-based filtering of PubMed papers improved verdict performance. Next, for health claims, PICO-based (Population, Intervention, Comparison, Outcome) paper mapping and summarization improves transparency of evidence used for verdict generation (like 'Berries reduce blood pressure'). Still, the RAG-LLM models we tested have biases towards positivity (too many foods deemed heart healthy) and neutrality (no clear direction). We discuss mechanisms at play and challenges on the route forward.

Keywords:
health self-management,
eHealth,
AI,
LLM,
RAG,
claim verification,
hypertension,
cardiovascular disease,
nutrition

University of Maribor Press

1 Introduction

Cardiovascular disease (CVD) is a leading cause of global mortality (Badimon, 2019; Gaidai, 2023). Hypertension, or high blood pressure, is a significant risk factor for CVD and the leading global cause of disease. Recent literature highlights the importance of evidence-based Health Self-Management (HSM) in improving cardiovascular health and reducing healthcare costs (Dineen-Griffin, 2019).

Literature on HSM and cardiac health is increasing rapidly (Qama, 2022). Information overload hinders timely access to insights useful in HSM support. As a special challenge in nutritional science, there is a general perception that studies are often contradictory (Nagler, 2014; Armitage, 2019). Moreover, fabricated science[1] and large food industry lobbies exist, resulting in fabricated guidelines[2]. Studies have found that nutrition confusion is associated with nutrition backlash. For example, nutrition backlash decreased engagement in fruit and vegetable consumption (Lee, 2018). Given these challenges, we focus on nutrition (foods) and scientific evidence on how they help (or not) improve cardiac health and hypertension.

Our previous research on information needs and sources found that it is difficult for patients and practitioners to find actionable lifestyle advice which incorporates state-of-the-art scientific evidence (Simons, 2021, 2022a, 2023a): the Top 3 Dutch health institutes (for either hypertension or type 2 diabetes) provided watered down and inconsistent health advise, whereas as Google Scholar search heuristics analysis showed that the returned papers drown people in details and nonactionable research. In 2024, we surveyed 'expert users' (with LLM experience and who had just completed an intensive hypertension improvement challenge; Simons, 2022b, 2023b, 2024a, 2025) for their information needs and perceived added value of LLM's to help summarize health literature findings. In summary, they expressed concern about LLM's output and usefulness regarding (Simons, 2024b):

[1] Dr Neal Barnard (2018) eloquently explains how claims on cardiac health of eggs have (incorrectly) become more positive in the past decades, exactly because the previous decades had been so exhaustive on the negative cardiac health effects. In short, 'serious research' moved elsewhere, leaving a void filled by the egg industry to fabricate recent studies & reviews with designs to 'prove' healthiness.
[2] Even in the US Dietary Guidelines Advisory Committee, where objectivity should be key, 19 out of 20 members have clear industry affiliations and conflicting interests (Mialon, 2022).

- LLM information quality (and hallucinations, Sallam, 2023, Raina, 2024),
- dealing with conflicting health claims,
- explaining why updated advice is distinct from traditional/familiar advice,
- correct links & transparency regarding original studies used.

Still, advances in Artificial Intelligence (AI) like explicit claim verification and Retrieval-Augmented Generation (RAG) may mitigate these risks. So, we developed and tested an enhanced prototype, to help answer the ***Research Question***:

To which extent can enhanced RAG-LLM models improve evidence inclusion and verdict transparency in mining nutrition science for cardiac health and hypertension claims?

2 Related work & Prototype design

The ***claim verification*** task is studied under the umbrella of automated fact-checking (Guo, 2022). The task involves automatically verifying the authenticity of claims based on the retrieval of evidence. A conventional framework of claim verification consist of three modules: the retrieval of relevant documents given a claim, selecting evidence from documents, and predicting a label (true, false, or not sure) based on the top-k evidence (Wadden, 2020; Pradeep, 2021; Soleimani, 2020). Research gap analyses (Gao, 2023; Wu, 2024; Liu, 2024) led us to focus on improving quality of evidence included and of transparency of verdict generation.

Liu et al. (2024) used the traditional three-step approach of claim verification but focused on an ***Retrieval Augmented Generation (RAG)*** module to specifically focus on RCT studies as evidence base for a given COVID claim. Instead of retrieving evidence from a prepared database, this augmented retrieval module presented a real life scientific use case. Below, we introduce Inclusion-Criteria-based filtering to enhance RAG by improving relevance of the inputs used. Retrieval-augmented methodologies harness the capabilities of multiple information retrieval techniques such as document vectorization, semantic similarity-based retrievers, and similarity ranking mechanisms.

We formulate concise claims using the ***Population, Intervention, Comparison, and Outcome (PICO) framework*** (Richardson, 1995). This framework is commonly used to formulate good clinical research questions, which can be utilized to formulate clinical claims by adapting the elements to suit the nature of the claim being made (Huang, 2006). For example Liu et al. (2024) used the PICO framework to construct a Covid Verification dataset of 15 PICO-encoded drug claims.

We introduce PICO use to translate ***nutrition*** science into health claims. PICO-encoded health claims can enhance document retrieval by guiding the search toward semantically relevant papers. Next, PICO-based paper summarization towards claim verdicts improves transparency. As an example PICO-based health claim:

- Population: Adults with high cholesterol levels
- Intervention: Consumption of flax seeds
- Comparison: Standard diet without flax seeds
- Outcome: Reduces LDL cholesterol levels
- Claim: *Consumption of flax seeds reduces LDL cholesterol in adults with high cholesterol.*

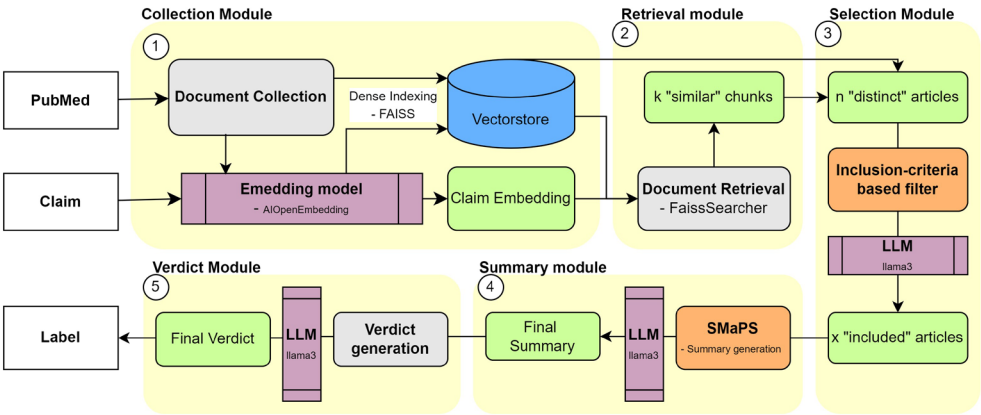

Figure 1: Framework with enhanced selection and summary modules

Figure 1 shows the overall 'Advanced RAG-LLM' framework we developed, including our extensions in orange. The numbers 1 to 5 highlight the key steps:

1. Document collection: based on semantic similarity, for each claim the most relevant articles are selected through the PubMed API. Each article (full text) is chunked into pieces of 1000 tokens to enable processing. FAISS [3] (Facebook AI Similarity Search) is used for indexing, similarity search and clustering of dense vectors, to store these vectors.
2. Retrieval: In our case, a health claim serves as input query and FAISS searches the vectorstore for the documents that are semantically similar to the query vector. However, it is crucial to understand that semantic similarity does not necessarily equate to relevance or quality, see step 3.
3. Selection: An Inclusion-Criteria based filter was added to increase relevance of the selected papers. The PICO elements were used: Population = human adults (e.g. not animals). Intervention = Dietary Intervention (e.g. not a prospective study, or a medication intervention). Comparison = Control group or -condition stated. Outcome = blood pressure or cardiovascular health (e.g. not bone density etc).
4. Summary: Using SMaPS (see Figure 2), PICO- and summary texts per article are summarized into a final summary for a given claim.
5. Verdict: Based on the final summary for a claim, a verdict is created, a score from 1 ('strongly refuted') to 5 ('strongly supported').

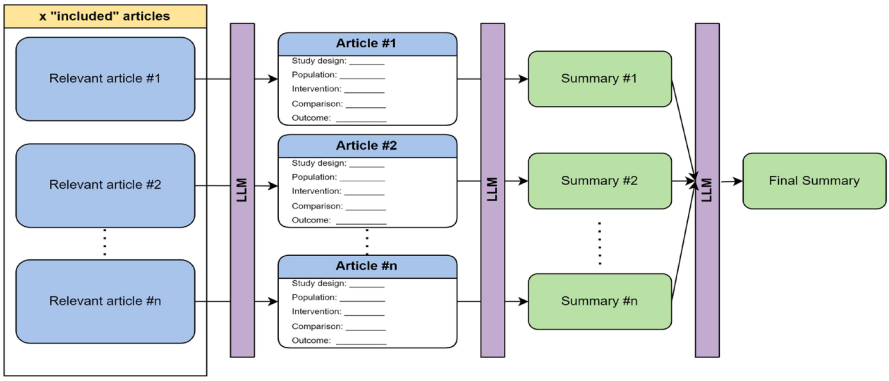

Figure 2: PICO-based Sequential Mapping (SMaPS)

[3] https://faiss.ai/

Figure 2 shows how the top k articles for a claim (e.g. 'Berries reduce blood pressure') are synthesized using LLM's towards PICO results and summaries, to create a final summary that can be used for verdict generation for that claim.

3 Method: Evaluation

Together with a cardiac health and nutrition expert, an initial set of 50 food and cardiac health claims was created (e.g. 'Legumes lower blood pressure in human adults'). Half of them focused on blood pressure, half of them on cardiovascular health. Next, from the 200,000 PubMed papers specifically on nutrition and cardiac health, 10,000 papers were selected, based on highest semantic similarity with the 50 claims. Given the fact that we were processing full texts of the papers, 10,000 papers was the maximum which was feasible with the computing resources we had available to develop and test our prototype.

Next, we evaluated the effectiveness of the 'Advanced RAG-LLM' prototype. The *three main subquestions for evaluation*:

1. Selection Module: How accurate is the Inclusion-Criteria-based Selection Module?
2. Summary Module: How accurate and useful are PICO- and summary-synthesis?
3. Verdict Module: How accurate are the verdicts of the 'Advanced RAG-LLM' prototype model, in comparison to expert opinion?

For subquestion 1, on accuracy of the Inclusion-Criteria-based Selection Module, we conducted a manual check for 100 articles which were predicted by the model as 'not' fitting and 100 articles as 'yes' fitting the Inclusion-Criteria.

For subquestion 2, on accuracy of PICO synthesis and usefulness of summaries, we conducted a first, low level 'expert user' evaluation of PICO- and summary texts for specific scientific articles: which were read by these experts as part of their evaluation. Given the time constraints in this prototype engineering study, we followed the approach of early-stage, iterative small scale user testing. Generally, in this approach the first 80% of design flaws appear with the first 5 user tests (Faulkner, 2003). For the evaluation, we used 5 healthy lifestyle experts who were

familiar with both interpreting scientific studies and with coachee information needs. Each of them evaluated 2 articles; 5 articles were evaluated, and each article was evaluated by 2 experts. Overall, 10 article evaluations were done and we observed a high degree of consistency across expert evaluations.

For subquestion 3, on accuracy of the food health verdicts of the 'Advanced RAG-LLM' prototype model, for all 50 claims, the model verdict was compared to an expert verdict from 1 ('strongly refuted') to 5 ('strongly supported'). The assignment for the expert was to generate verdicts following NutritionFacts[4] state-of-the-art. We compared three different model versions, see Figure 3: BaseLLM, Standard RAG-LLM, and Advanced RAG-LLM to highlight module differences. To evaluate model accuracy for all three models we used confusion matrices ('model predicted' versus 'expert' scores) and Cohen's weighted Kappa scores.

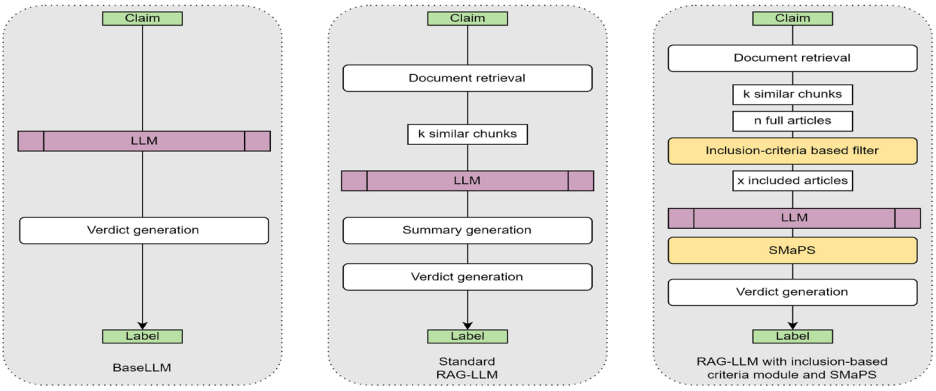

Figure 3: Comparing models: BaseLLM, RAG-LLM and Advanced RAG-LLM

4 Results

In response to ***subquestion 1, on accuracy of the Inclusion-Criteria-based Selection Module***, Figure 4 shows that the number of False Negatives (FN) are modest in comparison to True Negatives: 3 out of 100 model predictions. Compared to the actual n=83 papers for inclusion the 3 FN papers are 3,6%. False Positives (FP) however form quite a group: 20 (17.1%) of the actual n=117 papers for

[4] https://nutritionfacts.org/

exclusion are FP. Further analysis shows this is mostly due to incorrect flagging as intervention studies (of for example systemic reviews). Additionally, Population (animals) and Outcome (not really blood pressure) contributed to FP.

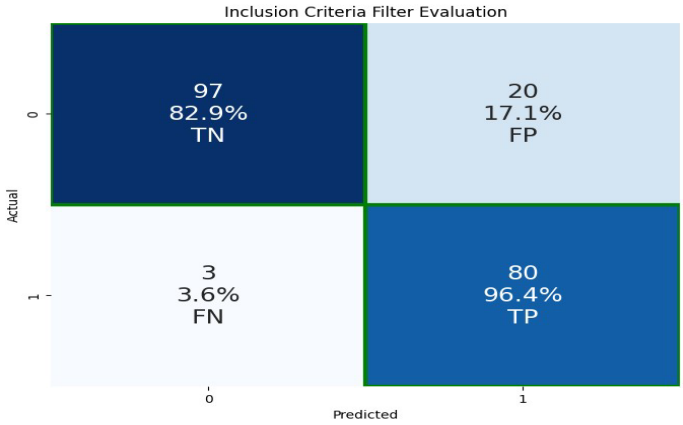

Figure 4: Confusion matrix for inclusion criteria: 0 = "excluded" and 1 = "included"

In response to ***subquestion 2, on accuracy of PICO synthesis and usefulness of summaries***, we found concerningly ***low performance***, see Figure 5. Whereas Intervention (4.2) and Comparison (4.1) score okay, accuracy scores for Population (2.0) and Outcome (1.9) descriptions were low. For Population accuracy in comparison to the original papers, some of the typical expert complaints were:

- "Missing info on sample size, selection method, cholesterol levels of study population."
- "Sample size incorrect, for intervention and for control. Selection method info absent."

For Outcome accuracy in comparison to the original papers, some of the typical expert complaints were:

- "Outcome incorrect; omission of the decrease in systolic blood pressure."
- "Primary outcome measures from paper were missing." → (10x out of 10)

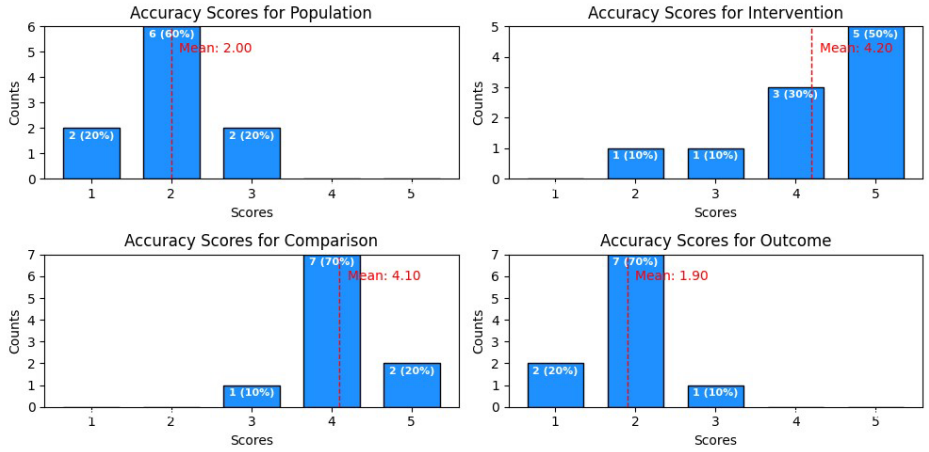

Figure 5: Expert Accuracy scores on PICO elements (n=10 paper evaluations)

Moreover, when ***evaluating usefulness of a paper's summary text***, the average expert score is 2.1 (on range of 1 to 5). Illustrative expert remarks:

- "Blood pressure is left out of the results. Which is strange because it's one of the main outcomes." [multiple times]
- "Claim cannot be made on this study. Study focusses on kiwi's only."

Regarding ***subquestion 3, on accuracy of the 50 food health verdicts of the three LLM models we compared***, we find that the Advanced RAG-LLM prototype generates improvements, but still LLM flaws persist, see Figures 6 and 7.

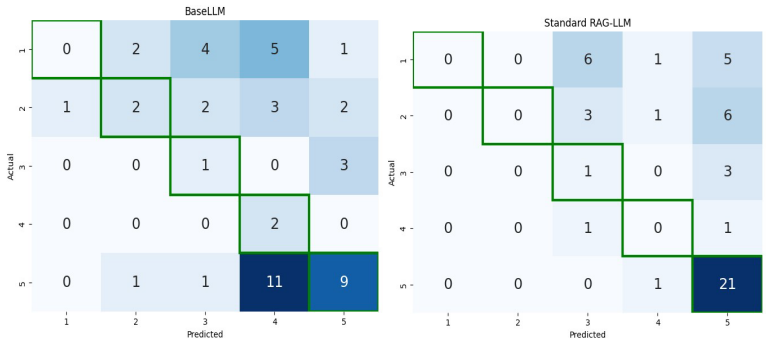

Figure 6: Confusion matrices BaseLLM & Standard RAG-LLM (n=50 claims predict/actual)

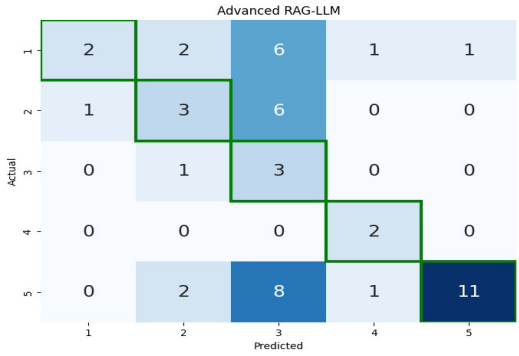

Figure 7: Confusion matrix Advanced RAG-LLM (n=50 claims: predict/actual)

The ***BaseLLM model has a 'positivity bias'***: most claims (foods) are predicted to be neutral (3) or healthy (4 or 5). The ***Standard RAG-LLM has even more 'positivity bias'***: Most claims (foods) are predicted to be clearly healthy (36 claims score 5; and 3 claims score 4). The rest (11 claims) are predicted as neutral: score 3. Interestingly, ***the Standard RAG-LLM model labelled no foods at all as unhealthy for blood pressure or cardiac health*** (not even sugar, alcohol, red meat, eggs, full fat dairy, cheese, salty foods, etc). As if there were not a single food detrimental to blood pressure or cardiovascular health in PubMed science...

Table 1: Model-Expert agreement: Cohen's Weighted Kappa

Model	Cohen's Weighted Kappa
BaseLLM	0.31
Standard RAG-LLM	0.27
Inclusion-based RAG-LLM	0.48

The ***Advanced RAG-LLM prototype generates more balance*** (positive and negative scores) and its verdicts are closer to expert opinion, see also its Kappa score of 0.48 in Table 1, meaning 'weak' agreement. Which is better than the 'minimal' agreements of 0.31 and 0.27 of the BaseLLM and Standard RAG-LLM models. Still, there is a bias towards neutral (3) scores and the Kappa scores indicate that there is significant room for improvement, which we discuss below.

5 Discussion & Conclusion

This prototyping study has several ***limitations***. Most importantly, due to limits in computing resources, only 10,000 paper full texts were used (5% of the 200,000 cardiac health nutrition papers present in PubMed: the 5% most semantically similar to the 50 claims used). We expect that including more papers will improve results, see also our discussion of verdict flaws below. Besides, the three evaluations conducted each have their limitations. First, the evaluation of Inclusion-Criteria based filter (subquestion 1) was a simple face value check. This could have been made more rigorous by adding more evaluators. Still, for humans these are simple evaluations (e.g. this is not a nutrition intervention on adults, but a meta study, an in vitro study or a mouse model study), so we expect low error rates here.

The second evaluation, expert-based assessment of paper PICO- and summary texts, was relatively small-scale (subquestion 2): 10 paper evaluations were done, by a total of 5 domain experts. Still, from a design perspective, the fact that for example in 10 out of 10 paper evaluations the Outcome text was found lacking to some extent in relevant information, is a sign to first improve this sequential summarization module before proceeding to larger-scale evaluations.

The third evaluation, for subquestion 3, on accuracy of the 50 food health verdicts of the three LLM models we compared, used the inputs of only one expert in translating the evidence-base from NutritionFacts[5] to verdicts. In a next phase, when the Advanced RAG-LLM is more robust, we can add more experts.

Still, the various user and expert evaluations do have face value: we can recognize the LLM model flaws that appear. Moreover, when digging deeper, we learn ***lessons*** why these specific flaws are the weaker points of model performance. For example, the ***'positivity bias'*** we found, especially in the BaseLLM and Standard RAG-LLM models, is interesting. We see two possible reasons. Firstly, standard LLM's (we used Meta's Llama3)[6] appear to be programmed to please the user (as various chess and other anecdotes in the history of LLM's have illustrated). But the second reason may be more substantial: in hindsight, all our health claims were formulated as positive

[5] https://nutritionfacts.org/
[6] We used Llama3 model from https://ollama.com/library/llama3 with a temperature setting of 0 to encourage the model to make prediction is purely based on the verdict with little creativity or variation.

statements (e.g. 'Berries reduce blood pressure'). In follow up research it would be interesting to see if different top k articles are selected when all claims are formulated as negative statements: '… does not reduce blood pressure').

Next, for some of the claims, not enough relevant intervention papers were included, which was exacerbated by including the wrong papers for a claim. This is illustrated by one of our explorative analyses: of the 50 claims, there were ***four claims (8%) with large differences (> 2 points) between expert- and Advanced RAG-LLM verdicts***. For these four claims (two claims on bananas and two on full fat dairy: claims for both blood pressure and cardiac health), we analysed the papers selected as evidence and a clear pattern appeared. After Inclusion-Criteria based filtering, only 3 to 5 papers were left as base material for these claims. Moreover, these were ***mostly 'false positive' papers***: included even though they were not specifically studying bananas or full-fat dairy. ***Thus, the LLM-based 'final summary' per claim was 'forced into' hallucination and overgeneralisation***. Though the prompting for study-referencing created transparent summaries (e.g. "The studies did not specifically investigate the effects of banana consumption on blood pressure." And "foods like fruits [..] support overall health") the evidence basis for a grounded verdict on the specific claim was absent (e.g. the n=5 studies included for the full fat and blood pressure claim were either focussing on low-fat dairy or dairy in general, but not on full-fat dairy).

Finally, ***one of the most pressing flaws we found, based on evaluating the PICO- and summary texts for papers, is the relatively poor representation of key elements/nuances from the scientific papers, as judged by the domain experts***. The outputs are lacking in nuance or context, rending paper summaries less useful (score 2.1, from 1 to 5). Since these poor representations are used as inputs for creating final summaries and verdicts per claim, these verdicts become 'averages of averages'. But in science, variance matters. For example, it makes a difference if we compare 0%-fat milk health effects to beans (0%-milk is comparatively 'unhealthy') or to twinkies (0%-milk is comparatively 'healthy'). Or if we make equal-weight or equal-calorie food swaps. So averaging averages is often not helpful, e.g. "If you want to know how fast a cheetah can run, taking its average speed of the day generates a very wrong number." This effect was aggravated by the fact that numbers (for study/control populations or outcomes) were often wrongly summarized or left

out. In short, the Llama 3 LLM and Facebook FAISS models we used appear to lack nuance for finding and summarizing the most relevant scientific facts.

Conclusion

The Advanced RAG-LLM prototype we tested does show improvements over base LLM and standard RAG-LLM performance. Firstly, we showed benefits from Inclusion-Criteria-based filtering. Secondly, we improved transparency with PICO-frame-based summarizing for verdict generation. Still, this could not disguise the fact that LLM's (= 'probable word generators') lack reasoning abilities and still disregard or misrepresent relevant facts from scientific papers. In short: The LLM-based models we tested are unable to distinguish relevant, solid science and findings from 'fabricated science' or low-relevance facts, thus they need extra reasoning tools.

Improved filtering and reasoning modules are needed to raise performance in:

- Paper selection
- Extracting key information (from: findings, control condition, population characteristics, study design, and expected causality of interventions)
- Context-based interpretation of numbers (e.g. 'Is this effect size large?')
- Highlighting findings which are most useful for patients and practitioners
- Recognizing and discarding 'fabricated science'

Acknowledgment

This research was (partly) funded by the https://www.hybrid-intelligence-centre.nl/ a 10-year programme funded the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022 and by EU H2020 ICT48 project ``Humane AI Net" under contract $\# $ 952026.

References

Armitage, H. (2019). Any way you slice it, there's a lot to say about nutrition studies. Scope blog Stanford, accessed 28-2-2025: https://scopeblog.stanford.edu/2019/01/28/any-way-you-slice-it-nutrition-studies-are-controversial/.

Badimon, L., Chagas, P., and Chiva-Blanch, G. (2019). Diet and Cardiovascular Disease: Effects of Foods and Nutrients in Classical and Emerging Cardiovascular Risk Factors. *Current Medicinal Chemistry*, 26(19):3639–3651.

Barnard R. J. (2018). How the Egg Industry Skews Science. YouTube, accessed 26-2-2024: https://www.youtube.com/watch?v=FyG8wr0gWIA.

Dineen-Griffin, S., Garcia-Cardenas, V., Williams, K., and Benrimoj, S. I. (2019). Helping patients help themselves: A systematic review of self-management support strategies in primary health care practice. *PLoS ONE*, 14(8), e0220116.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*, 379-383.

Gaidai, O., Cao, Y., and Loginov, S. (2023). Global Cardiovascular Diseases Death Rate Prediction. *Current Problems in Cardiology*, 48(5):101622.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. arXiv:2108.11896 [cs].

Huang, X., Lin, J., and Demner-Fushman, D. (2006). Evaluation of PICO as a knowledge representation for clinical questions. *AMIA Annual Symposium proceedings. AMIA Symposium*, 2006, 359–363.

Liu, H., Soroush, A., Nestor, J. G., Park, E., Idnay, B., Fang, Y., Pan, J., Liao, S., Bernard, M., Peng, Y., and Weng, C. (2024). Retrieval augmented scientific claim verification. *JAMIA Open*, 7(1), ooae021.

Lee, C.-j., Nagler, R. H., and Wang, N. (2018). Sourcespecific Exposure to Contradictory Nutrition Information: Documenting Prevalence and Effects on Adverse Cognitive and Behavioral Outcomes. *Health communication*, 33(4):453–461.

Mialon, M., Serodio, P., Crosbie, E., Teicholz, N., Naik, A., & Carriedo, A. (2022). Conflicts of interest for members of the US 2020 Dietary Guidelines Advisory Committee. *Public Health Nutrition*, 1-28.

Nagler, R. H. (2014). Adverse outcomes associated with media exposure to contradictory nutrition messages. *Journal of health communication*, 19(1):24–40.

Pradeep, R., Ma, X., Nogueira, R., and Lin, J. (2021). Scientific Claim Verification with VerT5erini. In Holderness, E., Jimeno Yepes, A., Lavelli, A., Minard, A.-L., Pustejovsky, J., and Rinaldi, F., editors, *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Qama, E., Rubinelli, S., and Diviani, N. (2022). Factors influencing the integration of self-management in daily life routines in chronic conditions: a scoping review of qualitative evidence. *BMJ Open*, 12(12):e066647.

Raina, A., Mishra, P., & Kumar, D. (2024). AI as a Medical Ally: Evaluating ChatGPT's Usage and Impact in Indian Healthcare. *arXiv preprint arXiv:2401.15605*.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12. Publisher: American College of Physicians.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare* MDPI, Vol. 11, No. 6, p. 887.

Simons, LPA, Neerincx MA, Jonker CM (2021). Health Literature Hybrid AI for Health Improvement; A Design Analysis for Diabetes & Hypertension, pp. 184-197, *34th Bled eConference*. June 27-30, Bled, Slovenia, Proceedings retrieval from www.bledconference.org. ISBN-13: 978-961-286-385-9, DOI: https://doi.org/10.18690/978-961-286-385-9 .

Simons, LPA, Neerincx MA, Jonker CM (2022a). Is Google Making us Smart? Health Self-Management for High Performance Employees & Organisations, *International Journal of Networking and Virtual Organisations,* Vol 27, No 3, pp.200-216. DOI: 10.1504/IJNVO.2022.10053605

Simons, LPA, Gerritsen, B, Wielaard, B, Neerincx MA (2022b). Health Self-Management Support with Microlearning to Improve Hypertension, pp. 511-524, *35th Bled eConference*. June 26-29, Bled, Slovenia, Proceedings retrieval from www.bledconference.org. ISBN-13: 978-961-286-616-7, DOI: 10.18690/um/fov.4.2022

Simons, LPA, (2023a). Health 2050: Faster Cure via Bioinformatics & Quantified Self; A Design Analysis, *International Journal of Networking and Virtual Organisations*, Vol 28, No 1, pp.36-52. **DOI**:https://doi.org/10.1504/IJNVO.2023.130957.

Simons, LPA, Gerritsen, B, Wielaard, B, Neerincx MA (2023b). Hypertension Self-Management Success in 2 weeks; 3 Pilot Studies, pp.19-34, *36th Bled eConference*. June 25-28, Bled, Slovenia, Proceedings retrieval www.bledconference.org. ISBN-13: 978-961-286-751-5, DOI: 10.18690/um.feri.6.2023

Simons, LPA, Gerritsen, B, Wielaard, B, Neerincx MA (2024a). Employee Hypertension Self-Management Support with Microlearning and Social Learning. *International Journal of Networking and Virtual Organisations*, *30*(4), 350-365. https://doi.org/10.1504/IJNVO.2024.140218.

Simons, LPA, Murukannaiah, PK, Neerincx, MA (2024b). Designing and Evaluating an LLM-based Health AI Research Assistant for Hypertension Self-Management; Using Health Claims Metadata Criteria, pp.283-298, *37th Bled eConference*. June 9-12, Bled, Slovenia, Proceedings. ISBN-13: 978-961-286-871-0, DOI: 10.18690/um.fov.4.2024.

Simons, LPA, Wielaard, B, & Neerincx, MA (2025). Beyond Average Results in Hypertension E-Support and Self-Management: Three Pilot Studies With Social Learning. In N. Wickramasinghe (Ed.), *Impact of Digital Solutions for Improved Healthcare Delivery* (pp. 257-282). IGI Global. https://doi.org/10.4018/979-8-3693-5237-3.ch009

Soleimani, A., Monz, C., and Worring, M. (2020). BERT for Evidence Retrieval and Claim Verification. In Jose, J. M., Yilmaz, E., Magalhaes, J., Castells, P., Ferro, N., Silva,˜ M. J., and Martins, F., editors, *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., ... & Xue, C. J. (2024). Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.