Can LLMs serve in identifying fake Health Information: it depends on how and who you ask.

Francois Bolduc, Ashwani singla, Manpreet Kaur, Mohammad Reza Taesiri, Keneizha Rubanarayana, Abhishek Dhankar, Osmar Zaiane, Marek Reformat

Submitted to: Journal of Medical Internet Research
on: October 08, 2024

Table of Contents

Can LLMs serve in identifying fake Health Information: it depends on how and who you ask.

Francois Bolduc[1] md, phd; Ashwani singla[1] msc; Manpreet Kaur[1] Msc; Mohammad Reza Taesiri[1] phd; Keneizha Rubanarayana[1]; Abhishek Dhankar[1] Mac; Osmar Zaiane[1]; Marek Reformat[1] PhD

[1]University of Alberta Edmonton CA

Corresponding Author:
Francois Bolduc md, phd
University of Alberta
3020 Katz Group Centre
Edmonton
CA

Abstract

Misleading information has significant implications for society but can have disastrous impact for health matters. Transformative artificial intelligence (AI) tools such as large language models (LLMs) have the potential for limitless content generation (including fake), soon making internet information impossible to assess using traditional human approaches. We asked if the same LLMs (GPT4 and Gemini1-5-Pro) could be part of a more scalable solution. We tested 2 publicly available LLMs for their ability to identify misinformation in HealthReleases previously labeled by human experts. We found that simple prompts lead to overall low accuracy (F1 Macro 0,45 (GPT4) and 0,49 (Gemini1-5Pro)), but very different profiles for each LLM. Adding specific criteria used by experts to critically assess the Releases enhanced Gemini (0.66) but surprisingly reduced GPT4 (0,37) performances. We therefore developed a novel approach incorporating summaries of expert feedback into prompts and then observed major improvements in performance for both LLMs(GPT4;0.63 and Gemini1-5Pro; 0.96). Our study provides the first use case of LLMs as high throughput proofing of medical text, but more importantly provides insights into LLMs' "truth biases". We provide a novel paradigm integrating knowledge into the prompts which may reduce the need for LLM training, and the requirement for ever larger datasets and compute power. Importantly, we show how experts could and need to be involved in LLMs used to enhance their performance and potentially minimize the data wall issue.

(JMIR Preprints 08/10/2024:67329)
DOI: https://doi.org/10.2196/preprints.67329

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Can LLMs serve in identifying fake Health Information: it depends on how and who youask.

AshwaniSingla[1],ManpreetKaur[1],MohammedRezaTaesiri[3],KeneizhaRubanarayana[1], Abhishek Dhankar[2], Osmar Zaiane[2], Marek Reformat[3,4], Francois V Bolduc[1,5,6]

1 Department of Pediatrics, University of Alberta, Edmonton, Alberta,Canada
2 Department of Computing Science, University of Alberta, Edmonton, Alberta,Canada
3 Electrical and Computer Engineering Department, University of Alberta,Canada
4 InformationTechnologyInstitute,UniversityofSocialScience,Poland
5 Neuroscience and Mental Health Institute, University ofAlberta,
6 WomenandChildrenHealthResearchInstitute,UniversityofAlbertaAbstract

Misleading information has significant implications for society but can have disastrous impact forhealthmatters.Transformativeartificialintelligence(AI)toolssuchaslargelanguagemodels (LLMs)havethepotentialforlimitlesscontentgeneration(includingfake),soonmakinginternet information impossible to assess using traditional human approaches.We asked if the same LLMs (GPT4andGemini1-5-Pro)couldbepartofamorescalablesolution.Wetested2publicly available LLMs for their ability to identify misinformation in HealthReleases previously labeled by human experts. We found that simple prompts lead to overall low accuracy (F1 Macro 0,45 (GPT4) and 0,49 (Gemini1-5Pro)), but very different profiles for each LLM. Adding specific criteria used by experts to critically assess the Releases enhanced Gemini (0.66) but surprisingly reduced GPT4 (0,37) performances. We therefore developed a novel approach incorporating summaries of expert feedback into prompts and then observed major improvements in performance for both LLMs(GPT4;0.63 and Gemini1-5Pro; 0.96). Our study provides the first use case of LLMs as high throughput proofing of medical text, but more importantly provides insights into LLMs' "truth biases". We provide a novel paradigm integrating knowledge into the prompts which may reduce the need for LLM training, and the requirement for ever larger datasetsandcomputepower.Importantly,weshowhowexpertscouldandneedtobeinvolvedin LLMs used to enhance their performance and potentially minimize the data wall issue.

Keywords:Largelanguagemodel,GPT4,Gemini1-5Pro,artificialintelligence,naturallanguage processing, LLM tuning, health, medicine, question answering, summarization, prompt engineering, news releases, misleading or fake news.

Introduction

Online platforms have greatly democratized access to medical information[1],but have also highlightedtheneedofevaluatingthequalityofinformationonunprecedentedscales,inorderto minimizet h e s h a r i n g o f m i s l e a d i n g i n f o r m a t i o n[2,3]. The rapid developments in generative artificial intelligence (AI) , especially large language models (LLMs), make scalable content evaluation even more needed, while leaving health experts and institutions to redefine not only how information is critically appraised but also how to engage with AI generated content. But, misinformation, especially when usinge x a g g e r a t i o n , s a t i r e , p a r o d y , f a b r i c a t i o n , m i s t a k e s[4]or emotionally charged messages[5]remains very challenging to detect for machines.

Nonetheless, there have been several attempts at using computers to detect misinformation. Pre-LLM,naturallanguageprocessing(NLP)[6-8]consistedoflinguisticfeatureextraction[9,10]combined with machine learning (ML) based multi-modal features-based misinformation detection[11,12],andevidence-basedfact-checkingwithdeeplearningalgorithms(BERT,SciBERT, and RoBERTa)[13,1415]. Using a dataset of news releases related to the health (and annotated by experts as true or misleading)[16], fine tuning[17]and inclusion of linguistic features[18]to the ML model allowed for detection of misleading information but with relatively low accuracy (0.616- 0.658 MacroF1), as F1 above 0.7 is generally accepted (REF). But, integrating user and content-relatedinformation(numberoffollowers,verifiedstatus,numberofretweetsandlikes) enhanced accuracy (0.843 MacroF1)[19]. While promising, these required dedicated teams, large computingpower,aredomainspecificandremainnotaccessibletomosthealthprofessionalsor the lay public.

LLMs have the potential to transform text generation[20]f o r m e d i c a l q u e r i e s , m e d i c a l e x a m i n a t i o n s , a n d m e d i c a l a s s i s t a n t s . [21–23]Combined with chatbots (as seen with OpenAI ChatGPT3[24]) LLM can be intuitively accessed by both the general public[2526]and health professionals.[27]Moreover, ChatGPT has recently been used for fact-checking general information.[28,29]On the other hand, LLMs have been shown to confabulate[30–32], harbor gender andracebiases[33,34],orrespondbasedonoutdated,unreliable,ordomain-nonspecifictrainingdata[35,36]. Moreover, LLMs' answers are influenced by the way the question (prompt) is designed.[37]38,39

Here we show that publicly available LLM-chatbots, GPT4[40]and Gemini can detect misleading information from health-related news releases. We also elucidate key aspects in prompt engineering, including a novel approach integrating expert input in order to improve performancesignificantly.O u r workshowshowintegratingexpertknowledgeinpromptscould mitigate the data wall and thus help reduce compute needs.

Method

Datasetusedformedicaltexts.Weleveragedthesamepublisheddataset[16]p r e v i o u s l y usedin other publications[17][18][19]i n c l u d i n g 6 0 6 t e x t s (HealthRelease) a n n o t a t e d b y d o m a i n e x p e r t s (journalism, medicine, health services research, and public h e a l t h) w i t h a g r o u n d t r u t h v a l u e (3 1 5 r a t e d b y e x p e r t s a s t r u e a n d 2 9 1 a s f a k e)a s w e l l a s i n d i v i d u a l l a b e l s o n 1 0 s t a n d a r d i z e d q u e s t i o n s (c r i t e r i a)[41](**Supp. Table 1**).1 4 r e s o u r c e s (n e w s _ r e v i e w s _ 0 0 0 1 8, 0 0 1 3 6, 0 0 1 6 8, 00235, 00284, 00395, 00422, 00490, 00499, 00507, 00589, 00590, 00601, 00605) out of 606are

not considered (no description text or no news_source wasprovided).

Prompt engineering. We first tested a zero-shot prompting approach: (**Prompt1**) where GPT4 or Gemini-1.5-Pro are instructed to detect misleading informationi n t h e N e w s R e l e a s e w i t h o u t p r o v i d i n g anycontext.Torestricttheoutput,itwasaskedtoshowonlythoseclaimswhichitcan provide a reference for (details in**Supp. Method file**).F o r t h e s e c o n d p r o m p t (**Prompt2**), we included in the prompt a specific criteria (quality of evidence, sensational language, novelty of approach, disease mongering and benefits of treatment, existing alternatives, harms of intervention, availability of treatment, cost of intervention, and conflict of interest) which was used also previously by experts to rate the NewsReleases[41](**Supp. Table 1**) as a context to make the model perform guided decisions along certain domain-specific dimensions. Additionally, the output instruction was modified to assign satisfactory or not satisfactory labels to each criteria andtheoveralltext.Forthethirdprompt(**Prompt3**),wefirstselected,foreachcriteriaanalyzed (quality of evidence, sensational language, novelty of approach, disease mongering and benefits of treatment), 50 NewsReleases labeled fake by experts and noted as not satisfactory for the criteria. We then asked GPT4 or Gemini-1.5-Pro to summarize into bullet point the justification provided by experts for marking a release as misleading (**Supp. Method File, Supp Table 2**).
Wethenincorporatedthesummarizationintotheprompt.Theremainingfivecriteria(existing

alternatives,harmsofintervention,availabilityoftreatment,costofintervention,andconflictof interest) were kept the same as prompt2.

Query process. We queried GPT4 and Gemini1-5-Pro using the requests library in Python following the request/response protocol. Initially, all the claim text description was stored in an ExcelsheetwhichwastakenasinputbythePythonscript.Alistofclaimtextswasiteratedanda prompt was created by embedding claim text descriptions into prompt templates to obtain the prompts for both experiment settings. Each prompt was passed as a message to Open AI or Gemini API calls along with other required API parameters with default values (role = user, frequency_penalty = 0.0, and temperature 0.0/0.2for GPT4 and default and temperature 0/1 for Gemini1-5Pro). The API returned a list of responses for a given prompt, and only the first response in the list was considered for further analysis. For prompt 3, we utilized LangChain framework[42]for GPT to create chat templates and the json output parser. Code is publicly available at GitHub[43].

Analysis: for prompt 1 and 2, GPT-4 and Gemini responses were further analyzed (using a Python script followed by manual result validation) to create a specific label (satisfactory or not satisfactory) for each output. For evaluating the prompts,the analysis of Identifying Fakes we used the following definitions: True Positive (TP): The Number of fakes identified correctly.False Positive (FP): The number of true items identified as fakes. True Negative (TN): The number of true items identified correctly as true. False Negative (FN): The number of fakes identified incorrectly as true. For the analysis of Identifying Trues: True Positive (TP): The number of true items identified correctly as true. False Positive (FP): The number of fakes identified incorrectlyastrue.TrueNegative(TN):Thenumberoffakesidentifiedcorrectly.False Negative (FN): The number of true items identified as fakes. We used those to calculate the Precision = TP / (TP + FP) and the Recall = TP / (TP + FN). The F1 score was calculated as before $2*((precision*recall)/(precision+recall))$.Macro average of the classification metrics(precision, recall and f1) is calculated by taking the unweighted mean of the metrics for each label using scikit-learn library[44].

Results

We started by asking if GPT4 or Gemini could label a NewsRelease as containing or not misleadinginformation(Prompt1,zeroshot)(**Fig.1A**).Importantly,thoseNewsReleaseswere labeled for ground truth by experts and used in ML based experiments for fake detection[16].
Importantlythedatasetcontainedbothtrue(315)andfake(291)resources.W e foundthatGPT-4 matchedexpertannotationfortruenewsin90.6%ofcasesbutwasonlyabletoidentify11.2%of the Fake ones, resulting in a low overall performance (as indicated by MacroF1 of 0.45)(**Fig. 1B-C;Supp. Table 3-4**). Surprisingly, Gemini1-5 Pro presented with an almost mirror image, lowperformanceforTrue(21.2%)andhighperformance(88.4%)forFake,leadingtoasimilar Macro F1 of 0.49 (**Fig. 1D-E;Supp. Table 3-5**). For both models, we found no significant differences by varying temperature from 0 to default settings (**Supp. Table 4-5**).

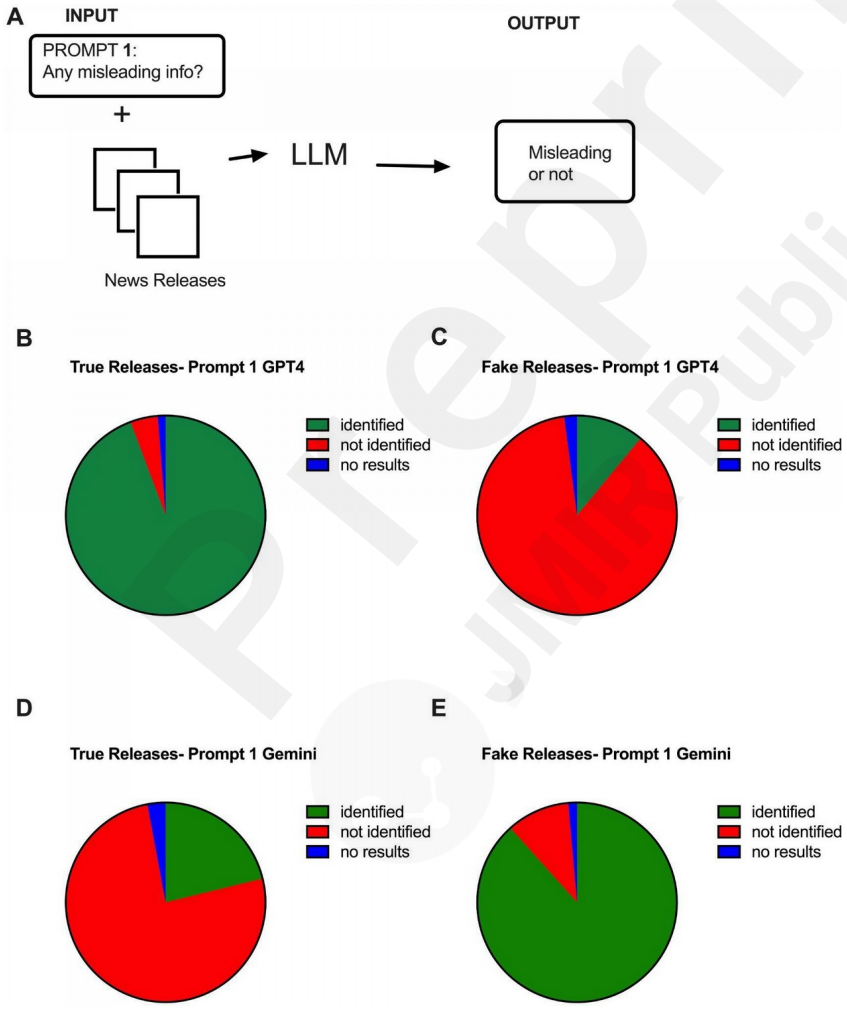

Figure 1. LLMs true/fake detection.**Prompt 1: Identify misleading information. A)**Schematic representation of the workflow. GPT4 or Gemini are asked to assess the NewsReleases for misleading information.**B)**For Releases labeled by experts as true, GPT4 identified90.6% as accurate, while reporting 4.2% of them as inaccurate.**C)**But, GPT4 only detected 11.2% of the Releases labeled as Fake by experts.**D)**In comparison, Gemini1-5Pro identifiedonly21.2%ofTrueReleasesassuchbutwasabletorecognizemisleadinginformation in 88.4% of Fake Releases.

Next, we postulated that using prompts similar to criterias used by experts to evaluate the NewsRelease would enhance LLMs performance (**Fig.2 A**). We therefore use the same criteria and ground truth scoring (more than 6/10 criteria scored as misleading is labeled Fake) to compareLLMandexpertlabels.Surprisingly,wefoundthatGPT4performanceincreasedforits ability to match experts on true news (97.5%) but worsened Fake detection (2.2%),resulting in low MacroF1(0.37)(**Fig. 2B-C); Supp. Table 3-4**). For Gemini1-5 Pro, we noticed a major improvement of performance for True (63.6%) but reduced for Fake (68.1%) nonetheless resulting in a more balanced performance and improved MacroF1 of 0.66(**Fig. 2D-E); Supp.**

Table3-5).

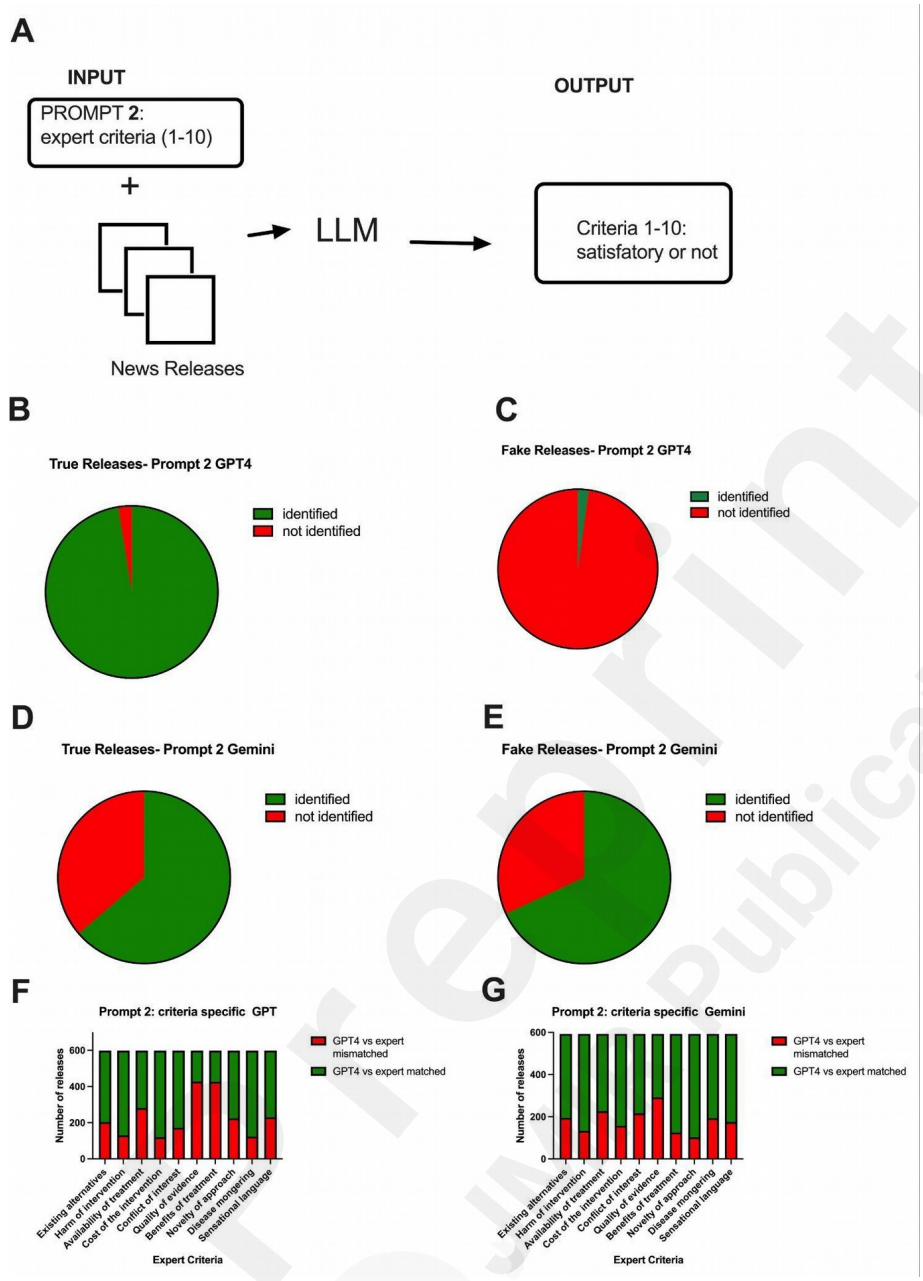

Figure 2. Criteria guided queries reveal uneven performance of LLMs. A) Schematic representation of the workflow. The LLM has to mark the release as satisfactory (or not) on specific criteria (asserted by experts on each release). 6 or more not satisfactory results in the Release being marked as Fake. GPT was able to identify **B) True** releases in 97.5% but **C) Fake in only 2.2%. On the other hand, Gemini identified D)** true in 63.6 % and **E)** fake in 68.1 %. Performance of **F) GPT4** and **G) Gemini** reveals heterogeneity in LLM responses based on individual criteria.

Examining performance on individual criterias revealed heterogeneous performance **(Fig. 2F-G)** in GPT4 and to a lesser extent in Gemini1-5Pro. Interestingly, examining justifications provided by both models (GPT4 for Fake and Gemini1-5Pro for True) revealed differences in bias toward critical thinking. For GPT4, when assessing disease mongering, experts raised the issue of the prevalence of conditions either lacking, not being supported by references, or cherry-picking the highest prevalence, when GPT4 thought the information provided was satisfactory. Also, controversies about disorders or treatments were not flagged by GPT4 but were noted by experts. Similarly, in releases fails related to the quality of evidence, GPT4 did not achieve enough critical analysis and did not pick on limitations (study design, power, data quality, outcome measures), methodological flaws (statistical analysis), or exaggeration of the significance of the results (translating findings, lack of peer review). In many cases, GPT4 marked the releases as true if the authors stated those limitations somewhere in the text instead of reporting the NewsRelease as misleading. Some issues in GPT4 understanding nuances about language became more explicit by looking at the question about Emotional or inflammatory language. Here, words like “breakthrough, cure, or revolutionary” were felt to be unsubstantiated by experts but marked as quotes and therefore not judged misleading by GPT4. Interestingly, GPT4 commented that those quotes were “optimistic or exciting” or justified considering the promising research instead of flagging them as inappropriate.

On the other hand, Gemini1.5-Pro struggled with True news by being very critical: reporting News as overly optimistic based on the evidence provided and reporting that authors were using laguage which was not conservative enough for health related topics.

So we hypothesized that by fine tuning the prompt (instead of the LLM itself) with input from the experts (through summarization), we could enhance (rebalance) performance for both LLMs (**Fig.3 A**). Iterative testing revealed that summarization of experts' feedback from misleading statements led to best performance (**Supp. Table 3**). For GPT4, prompt 3 boosted overall detection of fake releases to 43.4% but deflated the performance for true news (84.9%) **(Fig.3 B-C)** **(MacroF1** 0.63) (**Supp. Table 3-4**). For Gemini1-5Pro, we saw improvement in both True (97.4%) and Fake detection (93.7%) resulting in overall F1Macro 0.96 (**Supp. Table 3-5**).

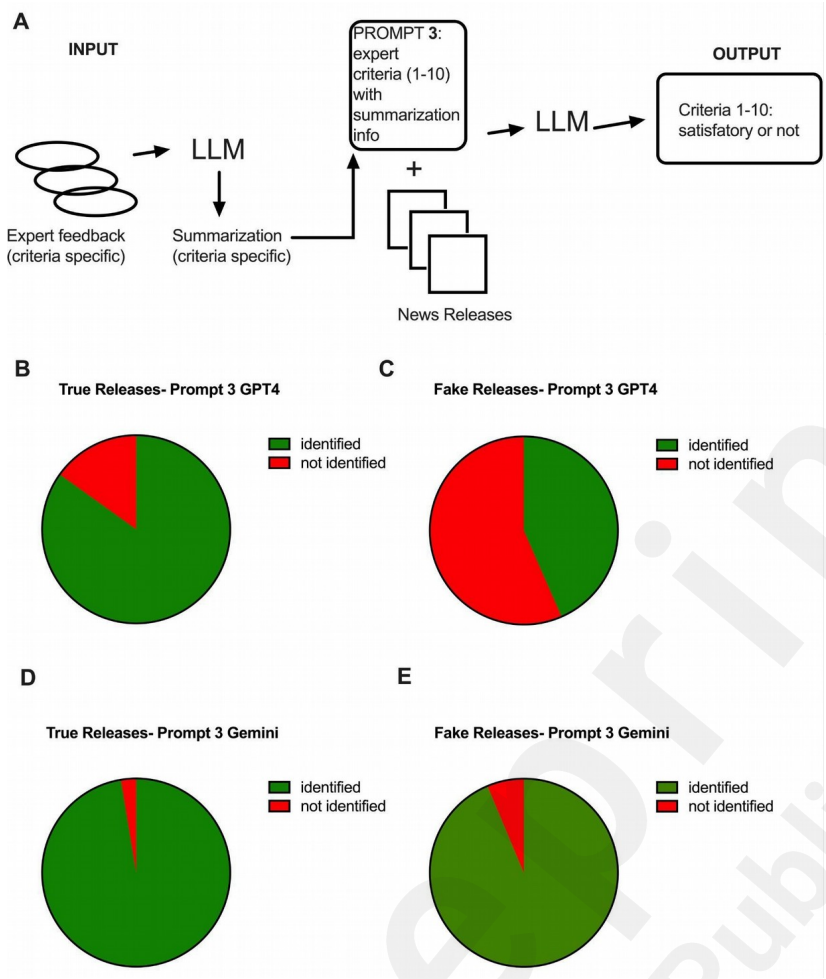

Figure 3. Integrating critical thinking into the prompt enhances LLMs performance. A) Schematic representation of workflow. The feedback provided by experts as to why they rated a NewsRelease as not satisfactory for an individual criteria (N=50 releases summarized per criteria) are summarized and integrated into the criteria specific prompt (3). GPT4 performance on **B)** True (84.9%) and **C)** Fake (43.4%) NewsReleases. Gemini1-5Pro performance on **D)** True (97.4%) and **E)** Fake (93.7%) releases.

Discussion

Finding ways of scaling the identification of misleading information remains challenging but is becoming a crucial aspect related to the value of online information. While we better appreciate now how misleading information ranging from “Fake News” to misinformation campaigns can beimpactfulonsocietyandpublichealth,wecanonlyimaginehowdisastrousitcouldbewhen itcomestohealth.Moreover,generativeAIopensatotallynewlevelofcapabilitieswithalmost unlimited voluntary (or not) machine based production of misleading information.

So, we asked if LLM could be used to evaluate information qualitybut also sought to better understand their inner workings. This is not trivial considering the limitations inherent to LLM: being predictive models, prone to confabulation, grown on different sets of data, released at different stagesofvalidationandrapidlyevolving.Recent workusingLLM(ChatGPT)toassert misinformationinpoliticalclaims[28,29,45,46],education[47,48],socialmediapostonCOVID[49,50]shows promising results but their usefulness for health related news remains untested.Using a dataset of News Releases labeled by experts as either True or Fake, we tested two publicly available LLMs (GPT4 and Gemini1-5Pro) for their ability at detecting misleading information. Not surprisingly considering recent work on prompt contextualization[51,52,39], the first zero shot approach led to relatively low accuracies but revealed how different LLMs would respond in very different ways: GPT4 labeled most articles as true, and therefore missed most of the Fake ones whereas Gemini1-5Pro proved overly critical and therefore labeled most true articles as containing misleading information. This could be due to differences in training data for each model, but recent work on another LLM, CLAUDE (Golden Gate version)[53]showed how individual entities could be assigned more weight and completely change the LLM’s output.

Another surprise was the relatively poor performance of both models when including specific scientific criteria as prompt in order to identify misleading information, showing that simply includingexpert-basedguidanceinpromptdevelopmentwasnotsufficientinachievingaccurate rating of the NewsRelease. This is important as it means that the inner knowledge of the LLM exert a major effect on its output, and would suggest that LLMs be grown on large amounts of data for complex applications such as science and health, adding further to the issue with compute needs/cost.

But, we found that injecting insights from experts on why they judged a NewsRelease was able forbothmodelto“rebalance”andenhancetheirperformancetolevelnotpreviouslyachievedby ML approaches (Gemini1-5 Pro MacroF1 0.96), giving us 1) a novel approach to promptdesign,
2) a novel approach for LLM fine tuning which would reduce compute needs andof the potential3) awayforwardforexpertstobeinvolvedinAIprogressesinsteadofsittingin margin. Our approach may help deal with the data wall

There are still several unresolved questions regarding the evolution of LLMs. For instance, their acceptability considering their rapid evolution, the relative opacity around the training data (including use of copyrighted material, and publicly created content) and our still basic understanding of their growth process. The emergence of on-device LLMs, while providing potentiallyenhancedprivacy,willalsomaketheevolvingnatureofLLMsanimportantaspecttostudy.

DataAvailability

WeutilizedtheopenlyavailableFakeHealthdatasetwhichcanbeaccessedat[54].Alltheanalysis scriptsanddatasetsusedforautomaticallylabelingtheGPT4responseareavailableinthepublic repository that can be accessed at.[55]

Ethicsdeclarations

The authors declare no competinginterests.

Authorcontributions:

AS and MK performed the experiments and co-wrote the manuscript. MRT performed the experiments and assisted with the manuscript. KR assisted with claim annotation. AD and OZ identifiedinitialresourcesforthedatasetasmisleading,contributedtotheproject'sconceptual development, and provided critical input into the manuscript. MR participated in design and

providedcriticalinputintothemanuscript.FVBconceptualizedtheproject,analyzedthedata, obtained funding, and co-wrote the manuscript.

Supplementary material

Supplementary Methodfile

SupplementaryTable1.Listofcriteriasusedbyexperts
SupplementaryTable2.SummariesGPT4forprompt3
Supplementary Table 3. Results metrics
Supplementary Table 4. Results GPT4
SupplementaryTable5.ResultsGemini1-5PRO

References

1. *PewResearchCenter.InternetHealthResources.PewResearchCenter:Internet,Science&Techhttps://www.pewresearch.org/internet/2003/07/16/internet-health-resources/*(2003).

2. Suarez-Lledo, V. & Alvarez-Galvez, J. Prevalence of Health Misinformation on Social Media: Systematic Review. *J. Med. Internet Res.* **23**, e17187 (2021).
3. *Swire-Thompson, B. Public Health and Online Misinformation: Challenges and Recommendations*. (2020).

4. Dhankar,A.,Zaïane,O.R.&Bolduc,F.UofA-TruthatFactify2022:TransformerAndTransfer Learning Based Multi-Modal Fact-Checking. (2022).
5. Ahmad,A.R.&Murad,H.R.TheImpactofSocialMediaonPanicDuringtheCOVID-19PandemicinIraqiKurdistan:OnlineQuestionnaireStudy.*J.Med.InternetRes.**22***,e19556(2020).
6. Sarrouti,M.,BenAbacha,A.,M'rabet,Y.&Demner-Fushman,D.Evidence-based *Fact-CheckingofHealth-relatedClaims.inFindingsoftheAssociationforComputationalLinguistics: EMNLP 2021*3499–3512 (2021).
7. Tan,N.*etal.Multi2Claim:*GeneratingScientificClaimsfromMulti-ChoiceQuestionsfor*ScientificFact-Checking.inProceedingsofthe17thConferenceoftheEuropeanChapterof the Association for Computational Linguistics*2652–2664 (2023).
8. Barve,Y.&Saini,J.R.Detectingandclassifyingonlinehealthmisinformationwith'Content Similarity Measure (CSM)' algorithm: an automated fact-checking-basedapproach.*J. Supercomput.***79**, 9127–9156 (2023).
9. Pritam Deka Queen's University Belfast, UK, Anna Jurek-Loughrey Queen's UniversityBelfast, UK & Deepak Queen's University Belfast, UK. Unsupervised KeywordCombinationQueryGenerationfromOnlineHealthRelatedContentforEvidence-BasedFact Checking.https://dl.acm.org/doi/10.1145/3487664.3487701doi:10.1145/3487664.348770 1.
10. Samuel, H. & Zaïane, O. MedFact: Towards Improving Veracity of Medical Information inSocialMediaUsingAppliedMachineLearning.*AdvancesinArtificialIntelligence108–120*(2018).
11. Silva,A., Luo,L., Karunasekera, S.& Leckie,C. Embracing DomainDifferences inFake

News:Cross-domainFakeNewsDetectionusingMulti-modalData.*AAAI**35***,557–565(2021).

12. Verschuuren,P.J.,Gao,J.,vanEeden,A.,Oikonomou,S.&Bandhakavi,A.LogicallyatFactify2:A Multi-ModalFactCheckingSystemBasedonEvidenceRetrievaltechniquesand Transformer Encoder Architecture. (2023).
13. Sarrouti,M.,BenAbacha,A.,M'rabet,Y.&Demner-Fushman,D.Evidence-based *Fact-CheckingofHealth-relatedClaims.inFindingsoftheAssociationforComputationalLinguistics: EMNLP 2021*3499–3512 (2021).
14. Xie,Q.*etal.Faithful*AIinMedicine:ASystematicReviewwithLargeLanguageModelsand Beyond.*medRxiv*doi:10.1101/2023.04.18.23288752.
15. ThestateofhumanNLPtechnologyforfact-checking.*Inf.Process.Manag.**60***,103219 (2023).
16. *FakeHealth:ThisRepository(FakeHealth)IsCollectedtoAddressChallengesinFakeHealth News Detection*. (Github).
17. *Debunkinghealthfakenewswithdomainspecificpre-trainedmodel.GlobalTransitionsProceedings***2**, 267–272 (2021).
18. Di Sotto, S. & Viviani, M. Health Misinformation Detection in the Social Web: AnOverviewandaDataScienceApproach.*Int.J.Environ.Res.PublicHealth**19***,2173(2022).
19. RishabhUpadhyayDepartmentofInformatics,Systems,andCommunication,UniversityofMilano -Bicocca, Italy, Gabriella Pasi Department of Informatics, Systems, andCommunication, University of Milano-Bicocca, Italy & Marco Viviani Department ofInformatics,Systems,andCommunication,UniversityofMilano-Bicocca,Italy.Leveraging

Socio-contextualInformationinBERTforFakeHealthNewsDetectioninSocialMedia.https://dl.acm.org/doi/10.1145/3599696.3612902doi:10.1145/3599696.3612902.

20. Ziyu,Z.*etal.Through*theLensofCoreCompetency:SurveyonEvaluationofLarge*Language Models. inProceedings of the 22nd Chinese National Conference onComputational Linguistics (Volume 2: Frontier Forum)*88–109 (2023).
21. Chang,Y.*etal.A*SurveyonEvaluationofLargeLanguageModels.(2023).
22. Clusmann,J.*etal.*Thefuture landscape oflarge languagemodels inmedicine. *Communications Medicine***3**, 1–8(2023).
23. Thirunavukarasu,A.J.*etal.Large*languagemodelsinmedicine.*Nat.Med.**29***,1930–1940(2023).
24. Panda, S. & Kaur, N. Exploring the viability of ChatGPT as an alternative to traditionalchatbotsystemsinlibraryandinformationcenters.*LibraryHiTechNews**40***,22–25(2023).
25. Xu,R.,Feng,Y.&Chen,H.ChatGPTvs.Google:AComparativeStudyofSearchPerformanc e and User Experience. (2023).
26. Marita Skjuve SINTEF Digital, Norway, Asbjørn Følstad SINTEF, N. & Petter BaeBrandtzaegSINTEF,NorwayandUniversityofOslo,Norway.TheUserExperienceofChatG PT: Findings from a Questionnaire Study of Early Users.https://dl.acm.org/doi/10.1145/3571884.3597144doi:10.1145/3571884.3597144.
27. Zakka, C.*et al.*Almanac: Retrieval-Augmented Language Models for ClinicalMedicine. *ResearchSquaredoi:*10.21203/rs.3.rs-2883198/v1.
28. *Hoes,E.,Altay,S.&Bermeo,J.LeveragingChatGPTforefficientfact-checking.PsyArXiv. April***3**,(2023).
29. Leite, J. A., Razuvayevskaya, O., Bontcheva, K. & Scarton, C. DetectingMisinformation

withLLM-PredictedCredibilitySignalsandWeakSupervision.(2023).

30. Emsley,R.ChatGPT:thesearenothallucinations–they'refabricationsandfalsifications. *Schizophrenia***9**,1–2(2023).
31. Alkaissi,H.&McFarlane,S.I.ArtificialHallucinationsinChatGPT:ImplicationsinScientific Writing.*Cureus***15**, e35179 (2023).
32. Azamfirei,R.,Kudchadkar,S.R.&Fackler,J.Largelanguagemodelsandtheperilsoftheir hallucinations.*Crit. Care***27**, 1–2 (2023).
33. Website.
34. Omiye,J.A.,Lester,J.C.,Spichak,S.,Rotemberg,V.&Daneshjou,R.Largelanguagemodels propagate race-based medicine.*npj Digital Medicine***6**, 1–4 (2023).
35. SummaryofChatGPT-Relatedresearchandperspectivetowardsthefutureoflargelanguage models.*Meta-Radiology***1**, 100017 (2023).
36. DeAngelis,L.*etal.ChatGPT*andtheriseoflargelanguagemodels:thenewAI-driveninfodemic threat in public health.*Front. Public Health***11**, 1166120 (2023).
37. *Giray,L.PromptEngineeringwithChatGPT:AGuideforAcademicWriters.Ann.Biomed. Eng.***51**, 2629–2633(2023).
38. Meskó,B.PromptEngineeringasanImportantEmergingSkillforMedicalProfessionals:Tutorial. *J. Med. Internet Res.***25**, e50638 (2023).
39. White,J.*etal.A*PromptPatternCatalogtoEnhancePromptEngineeringwithChatGPT.(2023).
40. Product.https://openai.com/product.
41. Dai,E.,Sun,Y.&Wang,S.GingerCannotCureCancer:BattlingFakeHealthNewswithaComprehensi ve Data Repository. (2020).

42. Introduction.https://python.langchain.com/v0.1/docs/get_started/introduction/.

43. Kaur,M.*MedicalFactChecker*.(Github).

44. *Classification_report.scikit-learn*

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html.

45. Satapara,S.,Mehta,P.,Ganguly,D.&Modha,S.FightingFirewithFire:AdversarialPrompting to Generate a Misinformation Detection Dataset. (2024).

46. Ni,J.*etal.AFaCTA:*AssistingtheAnnotationofFactualClaimDetectionwithReliableLLM Annotators. (2024).

47. O'Connor,S.*etal.Prompt*engineering whenusinggenerativeAIinnursingeducation. *Nurse Educ. Pract.***74**, 103825(2024).

48. Heston,T.F.*PromptEngineering:ForStudentsofMedicineandTheirTeachers*. (Independently Published, 2023).

49. Choi,E.C.&Ferrara,E.FACT-GPT:Fact-CheckingAugmentationviaClaimMatchingwith LLMs. (2024).

50. Xie,Q.*etal.Faithful*AIinMedicine:ASystematicReviewwithLargeLanguageModelsand Beyond.*medRxiv*doi:10.1101/2023.04.18.23288752.

51. Yao, Z., Cao, Y., Yang, Z., Deshpande, V. & Yu, H. Extracting Biomedical FactualKnowledgeUsingPretrainedLanguageModelandElectronicHealthRecordContext. *AMIA Annu. Symp. Proc.***2022**, 1188 (2022).

52. Wang,J.*etal.Prompt*EngineeringforHealthcare:MethodologiesandApplications.(2023).

53. GoldenGateClaude.https://www.anthropic.com/news/golden-gate-claude.

54. GitHub - EnyanDai/FakeHealth: This repository (FakeHealth) is collected toaddress

challenges in Fake Health News detection.GitHub

https://github.com/EnyanDai/FakeHealth.

55. *GitHub - ashwani227/gpt4AnalysisFakeHealth.GitHub*

https://github.com/ashwani227/gpt4AnalysisFakeHealth.

Supplementary Files

Multimedia Appendixes

Supplementary table 1.
URL: http://asset.jmir.pub/assets/220e56921d1a84550c4b5d95fafe8e77.xlsx

Supplementary Table 2.
URL: http://asset.jmir.pub/assets/3ca4786b273ab8505cbfcfa0bae06bd6.xlsx

Supplementary Table 3.
URL: http://asset.jmir.pub/assets/c8214c19a3a276af8017e8cac4219be2.pdf

Supplementary Table 4.
URL: http://asset.jmir.pub/assets/c4399230caaada2e0f5dc7df6a0c4279.xlsx

Supplementary table 5.
URL: http://asset.jmir.pub/assets/2afced551fa6e7fa2b2cc55846e88c0e.xlsx

Supplementary Method file.
URL: http://asset.jmir.pub/assets/9081f738ecc5b25519af6d05f22e1626.docx