Towards Automated End-to-End Health Misinformation Free Search with a Large Language Model

Ronak Pradeep(✉) and Jimmy Lin

David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada
{rpradeep,jimmylin}@uwaterloo.ca

Abstract. In the information age, health misinformation remains a notable challenge to public welfare. Integral to addressing this issue is the development of search systems adept at identifying and filtering out misleading content. This paper presents the automation of Vera, a state-of-the-art consumer health search system. While Vera can discern articles containing misinformation, it requires expert ground truth answers and rule-based reformulations. We introduce an answer prediction module that integrates GPT_x with Vera and a GPT-based query reformulator to yield high-quality stance reformulations and boost downstream retrieval effectiveness. Further, we find that chain-of-thought reasoning is paramount to higher effectiveness. When assessed in the TREC Health Misinformation Track of 2022, our systems surpassed all competitors, including human-in-the-loop configurations, underscoring their pivotal role in the evolution towards a health misinformation-free search landscape. We provide all code necessary to reproduce our results at https://github.com/castorini/pygaggle.

Keywords: Neural Retrieval · Large Language Models · Misinformation

1 Introduction

Individuals often resort to web search engines for acquiring health-related information, motivated either by curiosity or the need for self-diagnosis. However, the pervasive presence of false or misleading content complicates the distinction between accurate and inaccurate information or credible and non-credible sources. Misinformation can lead to reliance on ineffective and potentially harmful treatments, exacerbating health risks.

In recent years, the integration of information retrieval and deep learning has been instrumental in enhancing the accuracy and reliability of search results, highlighted by the advent of pretrained models like BERT [4]. These models have eclipsed traditional IR methods such as BM25 [9] in effectiveness. A multi-stage ranking approach has emerged, balancing model complexity and search latency by progressively narrowing candidate sets, facilitating the application of powerful yet slower rerankers [5].

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14611, pp. 78–86, 2024.
https://doi.org/10.1007/978-3-031-56066-8_9

Pradeep et al. [7] pinpointed the pitfalls of training multi-stage ranking models on datasets primarily consisting of credible information, leading to an unintended emphasis on harmful misinformation. In response, they introduced Vera, a stance prediction strategy effective at discerning useful from harmful content. Additionally, they advocated for the manual alignment of search questions to the appropriate stances to enhance result quality. This approach, however, is not without its flaws. The dependency on human-labeled stances and handcrafted rules for question reformulation restricts its adaptability and comprehensiveness, prompting a need for automation and generalization to accommodate the dynamic and diverse nature of information queries.

Through this paper, we attempt to tackle some drawbacks of Vera by automating the labeling of answers by organizers and generating arbitrary reformulations using large language models from the GPT class of models in combination with Vera [7].

More specifically, we build an answer prediction module using GPT_x in a zero-shot, in-context manner. We explore enhancing its capabilities by integrating it with evidence-aware Vera repurposed for the answer prediction task. Evaluating various prompting methods for GPT_x, including chain-of-thought reasoning, we develop a system that rivals human experts without requiring additional fine-tuning. Finally, we introduce a GPT_x reformulator that, taking into account these predicted answers, rephrases the question into a more naturally structured sentence, yielding improved retrieval effectiveness. On evaluating our systems in the TREC 2022 Health Misinformation Track, we find that our best systems outperform competing systems involving humans in the loop.

2 Datasets

2.1 MS MARCO Passage Ranking

We leverage relevance ranking models trained on the MS MARCO passage ranking dataset [1], which provides a corpus of 8.8M passages gathered from Bing search engine results. The collection has a training set of 500K (query, relevant passage) pairs, which helps finetune large language models, known to require a lot of training data to avoid overfitting the training data distribution.

2.2 TREC Health Misinformation

The TREC Health Misinformation Track is a concerted effort to advance retrieval methods that elevate accurate and credible health information while mitigating the spread of misinformation. In our study, we meticulously assess our systems utilizing the TREC 2022 Health Misinformation Track, drawing insights from questions and their associated medical consensus-based stances from the 2021 iteration to enrich our answer prediction module. Documents in this track are judged and categorized as “helpful” or “harmful”, contingent on the relevance grade assigned, leading to the compilation of two distinct judgment sets.

In evaluating retrieval effectiveness, organizers employed three primary metrics: helpful compatibility measure ($\text{COMP}_{\text{HELP}}$), harmful compatibility measure ($\text{COMP}_{\text{HARM}}$), and the difference between these measures (COMP_{Δ}) [2,3]. The objective is to amplify helpful content and demote harmful documents, benchmarked by the COMP_{Δ} metric.

The TREC 2022 Health Misinformation Track inaugurated the "Answer Prediction" task that requires teams to submit predicted answers for all topics. Unlike previous versions where the organizers supplied the medical consensus answer, this new requirement augments the complexity of the tasks and adds a layer to system evaluations. However, this addition is critical, given that end-to-end systems capable of automatically determining the medical consensus and leveraging them to improve retrieval effectiveness are crucial to dealing with the evolving nature of health information.

3 Multi-stage Relevance Ranking

3.1 First-Stage Retrieval

The first stage receives as input the user query, q, and produces top-k_0 candidates R_0 from the corpus. The intent is to curate a refined candidate set to be scored by a sophisticated neural reranker.

The query is treated as a "bag of words" for ranking documents from the corpus. We used a standard inverted index based on BM25 in the Anserini IR toolkit [12,13], built on the popular open-source Lucene search engine with default hyperparameters. In all experiments, we set $k_0 = 1000$.

3.2 Neural Rerankers

In this stage, documents retrieved by first stage retrieval, R_0, are reranked by a pointwise reranker called monoT5 [6]. The model estimates a score s_i, quantifying how relevant a candidate $d_i \in R_0$ is to a query q.

monoT5 uses T5 [8], a popular pretrained sequence-to-sequence transformer model. During training, the model takes in query–document pairs from MS MARCO passage [1] and produces the words "true" or "false" depending on whether the document is relevant to the query. Nogueira et al. [6] finetuned monoT5-3B (around 2.8B model parameters) with a constant learning rate of 10^{-3} for 10k iterations with a batch size of 128. Following [6], we reranked the documents according to the probabilities assigned to the "true" token.

We further refined our candidate set with duoT5-3B, which aims to predict $p_{i,j}$, the probability d_i is more relevant than d_j to the query. To do so efficiently, we used the representative segment of both d_i and d_j based on the highest monoT5 score. To account for the longer input length resulting from pairs of document segments, we increased the maximum number of input tokens from the default of 512 to 1024. At inference time, we aggregated the pairwise scores $p_{i,j}$ so that each document received a single score s_i.

Following Pradeep et al. [7], we evaluated pointwise and pairwise variants that used the topic description as the query (monoT5$_{base}$ and duoT5$_{base}$) and also those that rephrased the query based on the predicted answer into a natural language sentence form (monoT5$_{NL}$ and duoT5$_{NL}$).

```
We are a committee of leading scientific experts and medical doctors
reviewing the latest and highest quality of research from PubMED. For
each question, we have chosen an answer, either 'yes' or 'no', based on
our best understanding of current medical practice and literature.
Q: Will wearing an ankle brace help achilles tendonitis? E: There is
little scientific evidence to suggest that orthotics alone will be
effective in healing it. A: no
Q: Can vitamin b12 and sun exposure together help treat vitiligo? E: The
spread of vitiligo stopped in the majority of the patients after this
treatment. A: yes
...
Q: {query} E:
```

Fig. 1. The GPT$_x$ Prompt for the Answer Prediction module.

4 Stance and Answer Prediction Modules

In this section, *stance prediction* refers to the task where the model is given a question and a relevant text snippet and returns a label corresponding to the snippet's stance to the question. The *answer prediction* task is more global, where we aim to predict a single stance to a question that represents the medical consensus.

4.1 Vera—Stance Prediction

Pradeep et al. [7] addressed the problem of discerning misinformation by leveraging a stance prediction module called Vera. Given the topic q and a document d_i, the model is tasked to predict a label $\hat{y}(q, d_i) \in \{\text{true}, \text{weak}, \text{false}\}$. They leveraged the same input template as monoT5. To train Vera, they utilized effectiveness judgments from the TREC 2019 Decision (Medical Misinformation) Track and finetuned the Vera-3B model using a constant learning rate of 10^{-3} for 500 iterations with batch sizes of 128.

During inference, for a particular document d_i, given t_i and f_i are the probabilities assigned to the "true" and "false" tokens, respectively, they used the scoring scheme:

$$s_i^{\text{final}} = \lambda \cdot s_i^z + (1 - \lambda) \cdot \begin{cases} t_i - f_i, \text{answer field is "yes"} \\ f_i - t_i, \text{answer field is "no"} \end{cases} \quad (1)$$

which they denoted as Vera (λ, z), where $z \in \{\text{mono}, \text{duo}\}$ is referred to as the "relevance setting" and λ is the linear combination constant. The "weak" labels do not factor into inference as we are only concerned with how "true" or "'false" the model believes the stance is.

4.2 GPT_x—Answer Prediction

The success of Vera as a stance detection model relies on a single established medical consensus stance. These are absent (by choice) before judging in the 2022 edition. To this end, we formulate two ways to deal with this issue, the first of which leverages OpenAI's large language models, GPT_x.

We experimented with one completion model, GPT_3 (`text-davinci-002`) and two chat models, $\text{GPT}_{3.5}$ (`gpt-3.5-turbo`) and GPT_4 (`gpt-4`) using the prompt seen in Fig. 1. The prompt begins with a preamble of what we ideally strive for when we have access to labor and resources, a consensus among experts.

Providing such information in the prompt helps in grounding the model. Then, we included eight examples of questions from the TREC 2020 Health Misinformation Track, followed by an explanation leading to the answer inspired by chain-of-thought (CoT) reasoning [11].reasoning [11]. We handcrafted a short and simple explanation based on a quick skim of the "PubMED" article the decision makers cite. Finally, we added to the prompt a query from the TREC 2022 Health Misinformation Track and appended with the "E:" tag that signifies what comes next is the explanation. Note that this method does not include information on *any* documents in the prompt. Finally, we generated a single token, took the scores corresponding to the "yes" and "no" tokens, and normalized over these two scores when both were available.

We experimented with self-consistency checks [10] with multiple (5) target sequences but found the results are always consistent, especially with larger models. We crafted the prompt before submissions and stuck with it for post-hoc analysis and ablations, albeit introducing and removing components.

The costs of querying the GPT_3, $\text{GPT}_{3.5}$, and GPT_4 API are at most 0.002 USD, 0.001 USD, and 0.03 USD, respectively. Computationally, we can get answer predictions in less than a minute for the entire test query set, with $\text{GPT}_{3.5}$ being considerably faster.

4.3 Vera—Answer Prediction

To extend Vera [7] to the task of answer prediction, we first calculated the probabilities Vera assigns to the true and false tokens for the top 50 monoT5 documents. With these probabilities, we calculated the means of "true" and "false" scores over all documents and predicted the answer "yes" if that of the "true" token is higher and "no" otherwise. This approach essentially forms a consensus based on the top-retrieved documents; employing more effective retrieval systems could enhance the precision of answer predictions. Given the corpus-aware consensus from Vera and corpus-free prediction from GPT_x, we also evaluated the effectiveness of the mean prediction from Vera and GPT_x.

4.4 GPT_x Reformulation

Pradeep et al. [7] found that reformulating health questions to a natural language sentence based on the predicted answer results in better relevance ranking results. However, they handcrafted rules to solve this task, which does not generalize to arbitrary questions. To avoid this complication, we leveraged GPT_3, which can easily handle this natural language reformulation task. We used the following prompt: "Rephrase the questions to sentence-long answers based on the stance. Question: Will wearing an ankle brace help heal achilles tendonitis? Stance: no Answer: Wearing ankle brace does not help heal achilles tendonitis ...", with eight in-context examples followed by the question and stance from the topic set. Variants with and without the reformulator are denoted by $*_{\text{NL}}$ and $*_{\text{base}}$, respectively.

Table 1. Model Effectiveness on the TREC 2022 Health Misinformation Answer Prediction task.

Model	AUC	Accuracy	TPR	FPR
(a) Median	70.7	64.0	80.0	48.0
(b) Humans	94.0	94.0	88.0	0.0
(c) GPT_3	95.2	86.0	76.0	4.0
(d) Vera	82.1	68.0	84.0	48.0
(e) Hyb(GPT_3, Vera)	93.4	88.0	80.0	4.0
(f) $GPT_{3.5}$	86.0	86.0	80.0	8.0
(g) + CoT	94.0	94.0	88.0	0.0
(h) GPT_4 + CoT	94.0	94.0	96.0	8.0

5 Results

Table 1 reports the results on the Answer Prediction task from the TREC 2022 Health Misinformation Track. The reported metrics are Area Under the Curve (AUC), Accuracy, True Positive Rate (TPR), and False Positive Rate (FPR). For reference, row (a) provides the median TREC evaluation score, and row (b) provides the score from a human-in-the-loop submission. Rows (c)–(e) present our official submissions and (f)–(h) our post-hoc experiments.

Firstly, GPT_3 and Vera answer prediction models, rows (c)–(d), are more effective than the median. Among the submissions, GPT_3, row (c), has the highest AUC that demonstrates the effectiveness of these large language models even in a zero-shot setting. However, combining retrieval-based methods such as Vera with GPT_3, row (e), seems to improve the accuracy, the most critical for the retrieval task.

Moving from GPT_3 to the chat variant $GPT_{3.5}$ shows a similar accuracy but better TPR, row (f) vs. (c). Models post GPT_3 do not provide token probabilities, forcing us to set probabilities of selected tokens as a 1. Hence, we do not include hybrids with Vera, as they do not make sense anymore.

Finally, we find that chain-of-though prompting results in a considerable effectiveness boost, rows (g) vs. (f). Switching the chat variant to GPT_4, rows (h) vs. (g), does not seem to have a considerable effect.

Table 2 looks at the effectiveness of the retrieval task of the TREC 2022 Health Misinformation Track. For reference, row (a) provides the median score across all submissions in the track. Row (b) presents the second top-scoring submitted run, a manual submission. Rows (c)–(j) represent all our submitted runs, and rows (k)–(l) represent our post-hoc runs. Rows (j)–(l) consider progressively better answer prediction modules ending with an Oracle system.

With little surprise, pointwise reranking improves the helpful compatibility scores over that of BM25, rows (d) and (f) vs. (c). While pairwise rerankers generally improve over pointwise results, in this case, looking at rows (d) vs. (e) and (f) vs. (g), we see a surprising drop in effectiveness.

Table 2. Compatibility scores on the TREC 2022 Health Misinformation Ad Hoc Retrieval Task.

Retrieval Model	Answer Model	$COMP_{HELP}$	$COMP_{HARM}$	$COMP_{\Delta}$
(a) Median	-	0.2455	0.1465	0.0990
(b) WatS	Humans	0.287	0.140	0.147
(c) BM25	-	0.1928	0.1487	0.0441
(d) + $monoT5_{base}$	-	0.2838	0.1942	0.0896
(e) + $duoT5_{base}$	-	0.2780	0.1894	0.0886
(f) + $monoT5_{NL}$	GPT_3	0.3276	0.1264	0.2012
(g) + $duoT5_{NL}$	GPT_3	0.3216	0.1467	0.1749
(h) Vera ($\lambda = 0.0$)	GPT_3	0.2836	0.0971	0.1865
(i) Vera $(0.95, mono)$	GPT_3	0.3386	0.1168	0.2218
(j) Vera $(0.95, mono)$	Hyb(GPT_3, Vera)	0.3460	0.0894	0.2566
(k) Vera $(0.95, mono)$	GPT_4 + CoT	0.3528	0.0892	0.2636
(l) Vera $(0.95, mono)$	Oracle	0.3602	0.0797	0.2805

When introducing the answer prediction model module and the query reformulator, comparing rows (f) to (d) and (g) to (e), we notice it brings a considerable increase in helpful compatibility scores and a similar drop in harmful compatibility scores, as desired.

The introduction of Vera in the $\lambda = 0$ setting, i.e., label prediction alone, results in a model with worse helpful compatibility but better harmful compatibility score compared to $monoT5_{NL}$, row (h) vs. (f). Linear combinations with the neural relevance ranking system, as seen from row (i) onwards, bring an

effectiveness boost, finding a spot with better helpful compatibility but slightly worse harmful compatibility scores.

Improving the answer prediction model (based on accuracy) results in progressively better results, as evidenced in rows (i)–(l). Row (k) illustrates our most effective automatic system, noting a 79% relative improvement over competing systems by other teams based on the primary metric, COMP_{Δ}. Compared to a system with an oracle answer prediction module, i.e., with ground truth answer predictions, row (l), this system demonstrates comparable effectiveness showcasing its strength.

6 Conclusion

In this paper, we focus on automating the consumer health search pipeline—building an end-to-end system capable of discerning helpful from harmful health search results with experimentation in the TREC 2022 Health Misinformation Track. We build an effective answer prediction module using GPT_{x} in a zero-shot in-context fashion and augment it with the evidence-aware Vera. We explore various ways of prompting GPT_{x}, incorporating chain-of-though reasoning to build a system on par with human experts without finetuning. We incorporate a reformulator that takes in these predicted answers and rephrases the question in a natural language sentence, resulting in improved results. Coupled with a state-of-the-art retrieval misinformation-free consumer health search pipeline, our models outperform runs from other teams by over 79% based on COMP_{Δ}.

Acknowledgements. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. Bajaj, P., et al.: MS MARCO: a human generated MAchine Reading COmprehension dataset. arXiv arXiv:1611.09268 (2016)
2. Clarke, C.L.A., Smucker, M.D., Vtyurina, A.: Offline evaluation by maximum similarity to an ideal ranking. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM 2020, pp. 225–234 (2020)
3. Clarke, C.L.A., Vtyurina, A., Smucker, M.D.: Offline evaluation without gain. In: Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2020, pp. 185–192 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), NAACL 2019, pp. 4171–4186 (2019)
5. Nogueira, R., Cho, K.: Passage re-ranking with BERT. arXiv arXiv:1901.04085 (2019)
6. Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. In: Findings of the Association for Computational Linguistics, EMNLP 2020, pp. 708–718 (2020)

7. Pradeep, R., Ma, X., Nogueira, R., Lin, J.: Vera: prediction techniques for reducing harmful misinformation in consumer health search. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021, pp. 2066–2070 (2021)
8. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 1–67 (2020)
9. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference (TREC-3), pp. 109–126 (1994)
10. Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations, ICLR 2023 (2023)
11. Wei, J., et al.: Chain of thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, NeurIPS 2022 (2022)
12. Yang, P., Fang, H., Lin, J.: Anserini: enabling the use of Lucene for information retrieval research. In: Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017, pp. 1253–1256 (2017)
13. Yang, P., Fang, H., Lin, J.: Anserini: reproducible ranking baselines using Lucene. J. Data Inf. Qual. **10**(4), 16:1–16:20 (2018)