Fact-Checking Generative AI: Ontology-Driven Biological Graphs for Disease-Gene Link Verification

Ahmed Abdeen Hamed[1,2], Alessandro Crimi[1], Byung Suk Lee[3], and Magdalena M. Misiak[4]

[1] Sano Centre for Computational Medicine, Cracow, Poland
a.hamed@sanoscience.org
[2] Binghamton University, Complex Adaptive Systems and Computational Intelligence Laboratory, Binghamton NY, USA
[3] University of Vermont, Department of Computer Science, Burlington VT, USA
[4] Department of Physiology and Biophysics, Washington DC, USA

Abstract. Since the launch of various generative AI tools, scientists have been striving to evaluate their capabilities and contents, in the hope of establishing trust in their generative abilities. Regulations and guidelines are emerging to verify generated contents and identify novel uses. we aspire to demonstrate how ChatGPT claims are checked computationally using the rigor of network models. We aim to achieve fact-checking of the knowledge embedded in biological graphs that were contrived from ChatGPT contents at the aggregate level. We adopted a biological networks approach that enables the systematic interrogation of ChatGPT's linked entities. We designed an ontology-driven fact-checking algorithm that compares biological graphs constructed from approximately 200,000 PubMed abstracts with counterparts constructed from a dataset generated using the ChatGPT-3.5 Turbo model. In 10-samples of 250 randomly selected records a ChatGPT dataset of 1000 "simulated" articles , the fact-checking link accuracy ranged from 70% to 86%. This study demonstrated high accuracy of aggregate disease-gene links relationships found in ChatGPT-generated texts.

Keywords: ChatGPT, fact-checking, generative AI, biological graphs, biological ontology, network medicine

1 Introduction

The rise of new generative AI technologies holds both potential and concerns. Particularly, the emergence of ChatGPT [1] caused scientists to raise various concerns related to the capabilities and the inauthentic contents of such tools. Van Dis et al. identified five key priorities aimed at educating the general public about the potential of ChatGPT and formulating an effective response to this transformative AI tool. Among the five guidelines, fact-checking and human verification of ChatGPT contents were highlighted [2]. Inspired by such guidelines,

here we present our work on computational fact-checking of biological networks we constructed from ChatGPT-generated content. The utilization of biological ontology (i.e., Disease Ontology, Gene Ontology, Gene Ontology Annotations) give credibility to the biological terms that make up the nodes of the graphs. Using biological entities from ontology to extract and construct biological graphs from the biomedical literature offers trustworthy ground truth. Using network models and algorithms offer the rigor needed to perform fact-checking at the aggregate level. This study assumes a closed-world assumption [3–6], which sets the fact-checking scope within the knowledge embedded in the literature dataset and not beyond.

Knowledge graphs have been instrumental in advancing fact-checking methodologies, enabling structured and nuanced analyses of claims and assertions. For example, Tchechmedjiev et al. introduced ClaimsKG, a comprehensive knowledge graph that houses fact-checked claims, allowing informed queries on truth values and related aspects [7]. Vedula and Parthasarathy's work stood out by introducing FACE-KEG, a knowledge graph tailored to expound whether a statement is true or false, addressing the transparency gap in fact-checking [8]. Lin et al. made strides with ontology-based subgraph patterns, constructing graph fact-checking rules that integrate intricate patterns, capturing both topological and ontological constraints [9–11]. Notably, Ciampaglia et al. laid a foundation for fact-checking by leveraging knowledge graphs to scrutinize claims, drawing from reliable sources like Wikipedia [12].

Wang et al. harnessed entity category information, using prototype-based learning to enhance verification accuracy and reasoning capabilities in knowledge graph-based fact-checking, marking a significant advancement in this domain [13]. Khandelwal et al.'s approach encompassed structured and unstructured data from knowledge graphs to address the challenge of evaluating facts amidst growing data and misinformation [14]. Orthlieb et al.'s attention-based path ranking model exhibited promise in automating fact-checking through knowledge graphs, emphasizing interpretability and competitive results [15]. Another notable contribution came from Shi, who introduced ProjE, a neural network model that improved the completion of knowledge graphs and the accuracy of fact-checking [16].

Recent advancements further underpin the significance of knowledge graphs in fact-checking. The approach of Wang et al. leveraged category hierarchy and attribute relationships, showcasing the potential of knowledge structure in fact verification [17]. Amidst the COVID-19 outbreak, Mengoni's extended knowledge graph enabled enhanced claim validation through leveraging existing fact-checking reports [18]. Kim introduced weighted logical rules mining and evidential path identification in knowledge graphs, enhancing computational fact-checking [19, 20]. Zhu et al. designed a knowledge-enhanced fact-checking system, tapping into both unstructured document knowledge bases and structured graphs to robustly identify misinformation [21].

2   Methods

In this section, we present a comprehensive methodology for constructing a reliable knowledge framework to assess the quality of content generated using ChatGPT. Our approach is centered around the utilization of biological graphs as rigorous models that offer quantitative analysis of objective outcomes. Graphs as a tool is also being investigated for the advancement of Large Language Models (LLMs) [22] and ChatGPT technologies [23].

The proposed approach consists of six key steps which as a whole contribute to verification of the authenticity and accuracy of AI-generated biomedical text: (1) ChatGPT prompt-engineering and simulated-articles generation, (2) partial-match ontology term chunking to increase the recall of term matching, (3) ontology feature extraction, where partial terms are used as the means to feature identification in the literature and ChatGPT text, (4) proximity-based biological graphs construction for capturing the strongest links among the biological terms, (5) biological graph topological analysis, by analyzing the structural properties of each type and comparing them accordingly, and (6) algorithmic fact-checking to assert the facts.

2.1   Prompt-Engineering ChatGPT for Simulated-Articles Generation

Using the ChatGPT APIs, we engineered a prompt that has two roles: (1) the system role which is to command the ChatGPT engine to generate biomedical abstracts and (2) a user role which is to command ChatGPT explicitly to perform the task shown in Algorithm 1, repeating it as needed until a dataset of the desired size is produced.

Algorithm 1 ChatGPT Prompt Engineering for Article Generation

---

Require: The number n of simulated articles.
Require: The number w of words in each article.
 1: Generate a list of n simulated PubMed-style abstracts.
 2: For each abstract containing three fields: GPT-ID, Title, and Abstract, make it w words.
 3: Make the GPT-ID random, containing at most five letters and numbers.
 4: Return the abstracts in a valid JSON format as an array of JSON records.
 5: Investigate the biology of human disease-gene associations.
 6: Provide details related to diseases, genes, cells, organisms, and any FDA-approved drugs, and state any relationships.

2.2   Feature Extraction and Biological Graph Construction

Ontology terms are inherently detailed and lengthy. In biomedical literature, these long names are frequently abbreviated for convenience. For instance, the

term "female breast cancer" is often referred to as "breast cancer" in the text. Importantly, we maintained a connection between these bigrams and their corresponding original term IDs in the ontology, while also tracking their positions. Constructing the knowledge graphs required the following steps: (1) feature extraction using the diseases and gene ontology and (2) establishing the links among the terms extracted. The process of ontology feature extraction from text records is as follow: (1) it takes as in put a collection of abstract texts and an ontology containing terms, (2) reads each textual record in the collection to identifies mentions of ontology terms (and related bigrams) within the text, (3) checks if the term appears in the text. If the term is a single word, the algorithm records the position of the match, the term itself, and other relevant information. For terms with more than one word, the algorithm generates bigrams (pairs of adjacent words) and checks for their presence in the text. If found, it records the position, term, bigram, and additional information. The process terminates by producing a set of matches for each record, indicating where ontology terms and bigrams were found within the text. Concretely, we constructed two different undirected but weighted graphs of disease and genes nodes. The first type one was constructed publication-driven from the mentions of disease and genes occurring in a dataset of biomedical abstracts extracted from PubMed Central [24]. A disease and a gene are connected if they occur in the same abstract. Then the link is weighted with the distance among the terms. Both gene and diseases names are ontology terms from the Human Disease Ontology (DOID) [25, 26], and the Gene Ontology and Annotation (GOA) [27–29].

2.3   Fact-Checking ChatGPT Biological Graphs

The purpose of this step is to investigate the authenticity of contents gathered from ChatGPT and other generative AI models, and to test whether such contents may bridge the disease–genes gap in our understanding. In this regard, we propose a computational approach that captures how much true knowledge is stated in ChatGPT graphs and also identifies what may be considered noise or novelties. The idea is to compare the various link types (disease-gene, gene-gene, disease-disease) and determine how much they overlap with those in the ground truth literature graph. This offers fact checking at an aggregate level without having to verify the link semantics. Specifically, from 10 graphs constructed earlier, we implemented a process that systematically computes the number of edges in a ChatGPT-generated graph that coincide with edges in the corresponding graph derived from literature abstracts. While being in the search space, the algorithm also tracks the link to discern each type and evaluates the balance in the facts founds. It extracts all links before it also processes one link at a time, and checks it against the ground-truth graph constructed from the literature.

3   Results

We used various network metrics that compare the ChatGPT graphs with literature Graphs objectively. Table 1 encapsulates the essential metrics pertaining to

each type of the knowledge graphs (i.e., literature and ChatGPT counterparts).

Table 1: The statistical result of comparing 10 GPT graphs with 10 literature graphs generated from the same number of records.

Source	Metric	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
GPT	No of Nodes	70	80	80	86	80	74	66	75	75	79
PubMed		137	63	100	104	116	118	101	113	95	154
GPT	No. of Edges	110	138	113	141	120	116	108	116	124	139
PubMed		297	124	165	214	251	240	207	366	147	393
GPT	N/E Ratio	0.64	0.58	0.71	0.61	0.67	0.64	0.61	0.65	0.60	0.57
PubMed		0.46	0.51	0.61	0.49	0.46	0.49	0.49	0.31	0.65	0.39
GPT	No. of Diseases	54	64	65	67	60	57	54	60	59	61
PubMed		117	54	87	88	97	101	79	97	75	131
GPT	No. of Genes	16	16	15	19	20	17	12	15	16	18
PubMed		20	9	13	16	19	17	22	16	20	23
GPT	Gene-Gene Link No.	46	67	59	64	51	48	54	56	56	62
PubMed		229	86	120	151	196	192	148	311	94	316
GPT	Disease-Gene Link No.	54	50	45	58	50	47	47	48	49	58
PubMed		57	34	40	55	45	36	44	43	42	61
GPT	Disease-Disease Link No.	10	19	9	19	19	20	6	11	19	19
PubMed		11	4	5	5	10	10	14	10	11	16

Each consecutive two rows embody a distinct scenario for a given statistic, while the columns reference the dataset selected randomly by a given seed. The "No. of Nodes" 2-row denotes the count of all nodes, which symbolize diseases, genes. The "No. of Edges" 2-row unit quantifies the interconnections between nodes, reflecting relationships (e.g., disease - gene) or interactions (protein - protein). The "N/E Ratio" 2-row unit computes the balance between nodes and edges, potentially demonstrating the network complexity of each graph. The "No. of Diseases" 2-row enumerates disease-related nodes, while "No. of Genes" 2-row unit does the same for genes. The "No. of Disease-Gene Links" 2-row unit indicates associations between diseases and genes. "No. of Disease-Disease Links" underscores connections between different diseases. Lastly, "Number of Gene-Gene Links" 2-row unit unveils interactions among gene nodes. Collectively, this table provides an intricate glimpse into the network's composition, connectivity, and relationships within the biological and medical framework, fostering a deeper understanding of its underlying dynamics of each type. Figures 1a and 1a demonstrate the comparisons of nodes and edges between ChatGPT and literature, respectively.

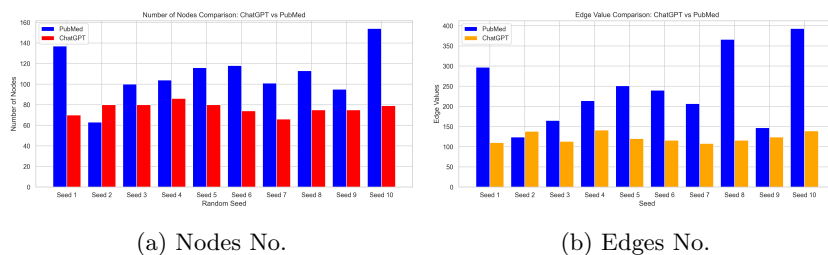(a) Nodes No.   (b) Edges No.

Fig. 1: shows two subfigures: (a) on the left, the Number of Nodes, and (b) on the right the Number of Edges comparisons of 10 chatGPT graphs against literature, respectively.

4 Discussion

The discussion of our study results revolves around several key observations and findings that shed light on the comparison and potential utility of the constructed knowledge graphs. Our approach involved the comparison of two distinct types of graphs, both constructed from randomly selected datasets. This sampling strategy ensured an unbiased and fair basis for comparison between the two sources. In terms of topological analysis, it was our expectation to observe a less number of nodes and edges exhibited in ChatGPT. It was also our expectations to observe that the network generated from literature to be rich and complex, which was demonstrated by lower ratios of nodes to edges. However, we also observed an anomalous behavior among the 10 graph. Particularly, G9 has surpassed its literature counterpart in the ratio of number of nodes to edges. Such an observation indicate complexity of certain ChatGPT graphs which warrant further pursuing.

One of the main pursuits of this work was to perform an unbiased fact-checking and verification of a truth graph constructed constructed using ontologies for their credible terminology, and biomedical literature of publications that are funded by the National Institute of Health to ensure high quality and credibility of work. We ensure that the fact-checking process is bounded by a closed-world assumption to make our work possible. The outcome of the process yielded promising results: the precision of link overlaps ranged from [70% to 86%] which is significantly high given the close-world assumption. This finding gives a certain measure of confidence to cautiously consider investigating data generated by ChatGPT using careful prompt-engineering.

Conclusion and Future Direction

As we continue to refine our work, the next steps involve further investigation of the proximity distance among biomedical terms and test if they hold in other domain and research areas. The study of disease-gene can be further instantiated in precise complex disease such as Alzheimer's and comorbidities where

little is known. Such investigations may necessitate the introduction of new ontologies (e.g., Gene ontology, Drug, Chemical Entity, and drug target ontologies) among many others. In turn, this opens the door to prompt-engineer ChatGPT to answer specific questions regarding the repurposeability of a drug. Another interesting direction is to entirely *retrain* the engines of ChatGPT using the confirmed-true knowledge and use its massive reasoning capabilities to answers questions about certain biological pathways to investigate a certain biological targets, or a disease that maybe caused by a certain clusters of genes.

5 Acknowledgments

References

1. OpenAI. ChatGPT: Conversational ai assistant. OpenAI Platform, 2023. Accessed: August 14, 2023.
2. Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.
3. Teodor C. Przymusinski. An algorithm to compute circumscription. *Artificial Intelligence*, 38, 1989.
4. Query rewriting for ontology-mediated conditional answers. 2020.
5. Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. 2011.
6. Jack Minker. On indefinite databases and the closed world assumption. volume 138 LNCS, 1982.
7. Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. Claimskg: A knowledge graph of fact-checked claims. In *The Semantic Web – ISWC 2019*, volume 11779, 2019.
8. Nikhita Vedula and Srinivasan Parthasarathy. Face-keg: Fact checking explained using knowledge graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 526–534, 2021.
9. Peng Lin, Qi Song, Jialiang Shen, and Yinghui Wu. Discovering graph patterns for fact checking in knowledge graphs. In *International Conference on Database Systems for Advanced Applications*, pages 783–801. Springer, 2018.
10. Panpan Lin, Quan Song, and Yanhong Wu. Fact checking in knowledge graphs with ontological subgraph patterns. *Data Science and Engineering*, 3:341–358, 2018.
11. Peng Lin, Qi Song, Yinghui Wu, and Jiaxing Pi. Discovering patterns for fact checking in knowledge graphs. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27, 2019.

12. Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.
13. Shuai Wang, Lei Wang, and Wenji Mao. A kg-based enhancement framework for fact checking using category information. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE, 2020.
14. Saransh Khandelwal and Dhananjay Kumar. Computational fact validation from knowledge graph using structured and unstructured information. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 204–208, 2020.
15. Téo Orthlieb, Hamdi Ben Abdessalem, and Claude Frasson. Checking method for fake news to avoid the twitter effect. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*, pages 68–72. Springer, 2021.
16. Baoxu Shi and Tim Weninger. Proje: Embedding projection for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'17, page 1236–1242. AAAI Press, 2017.
17. Shuai Wang, Wenji Mao, Penghui Wei, and Daniel D. Zeng. Knowledge structure driven prototype learning and verification for fact checking. *Knowledge-Based Systems*, 238, 2022.
18. Paolo Mengoni and Jinyu Yang. Empowering covid-19 fact-checking with extended knowledge graphs. In *International Conference on Computational Science and Its Applications*, pages 138–150. Springer, 2022.
19. Jiseong Kim and Key Sun Choi. Unsupervised fact checking by counter-weighted positive and negative evidential paths in a knowledge graph. In *Proceedings of the 28th international conference on computational linguistics*, 2020.
20. Ji-Seong Kim and Key-Sun Choi. Fact checking in knowledge graphs by logical consistency. *Semantic Web Journal*, swj2721, 2021.
21. Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29, 2021.
22. Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap, 2023.
23. Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling, 2023.
24. Pubmed central (pmc). Accessed: September 2nd, 2023.
25. Philipp Hofer, Sabrina Neururer, and Georg Goebel. Semi-automated annotation of biobank data using standard medical terminologies in a graph database. volume 228, 2017.
26. Adama Sow, Abdoulaye Guissé, and Oumar Niang. Enrichment of medical ontologies from textual clinical reports: Towards improving linking human diseases and signs. volume 296, 2019.
27. Rachael P. Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, and Claire O'Donovan. The goa database: Gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43, 2015.
28. Gene ontology annotations and resources. *Nucleic Acids Research*, 41, 2013.
29. Evelyn Camon, Michele Magrane, Daniel Barrell, David Binns, Wolfgang Fleischmann, Paul Kersey, Nicola Mulder, Tom Oinn, John Maslen, Anthony Cox, and Rolf Apweiler. The gene ontology annotation (goa) project: Implementation of go in swiss-prot, trembl, and interpro, 2003.