Harnessing the Power of Knowledge Graphs to Enhance LLM Explainability in the BioMedical Domain

Amir Hassan Shariatmadari [1], Sikun Guo [1], Sneha Srinivasan [1], and Aidong Zhang[1]

[1]Department of Computer Science, University of Virginia,
85 Engineer's Way, Charlottesville, VA 22903
ahs5ce@virginia.edu, qkm6sq@virginia.edu, gjf3sa@virginia.edu, aidong@virginia.edu

Abstract

This paper discusses the critical issue of enhancing the explainability and performance of Large Language Models (LLMs) in the biomedical domain by leveraging the structural benefits of Knowledge Graphs (KGs). Despite their impressive performance, LLMs often suffer from generating non-factual content, a phenomenon known as "hallucination". The lack of explainability exacerbates this issue, making it challenging to trust and interpret LLM outputs. By integrating KGs, which provides structured, interpretable data, we explore an approach to not only mitigate hallucinations but also to improve the transparency of the model's reasoning process. This paper demonstrates promising results when combining LLMs and KGs using the attention mechanism. Our method demonstrates plausible local explanations that illustrate reasoning between the LLM and KG while improving the performance of a biomedical reasoning task.

Introduction

Large Language Models (LLMs) bring a lot of potential to the biomedical domain. LLMs can be used to generate text, analyze biomedical data, personalize medical treatments, and much more. However, there are two key challenges inherent to LLMs that need to be addressed before they can be safely and effectively deployed for real-world use in a biomedical setting. While the state-of-the-art LLMs learn from a large corpus of data and perform well on evaluation datasets, they are prone to hallucinate; that is, they generate predictions or results that are non-factual or misleading in a confident manner. The ability to explain the model's predictions is vital to combat hallucinations. Explainability is also a challenge. The size and complexity [18] of LLMs make it difficult to have local and global self-intrinsic explanations.

To increase accuracy on downstream tasks, we look at fusing LLMs with knowledge graphs (KGs), since previous works combining the two demonstrate an increase in performance on various downstream biomedical tasks [28, 2, 22]. To increase model transparency for individual predictions, exploring attention is a good candidate for exploration, however, it does not come without controversy. Some researchers challenge the validity of attention as an explanation through various experiments [13, 19]. Others challenge this by demonstrating different ways attention can be explainable [1, 19, 20, 8].

In this work, we fuse LLMs with a biomedical KG and address the increased plausibility in attention-based explanations by directly attending LLM embeddings over KG embeddings. To demonstrate our contributions, we focus on biomedical reasoning with the MedQA dataset and leverage the fusion of BioBERT [16] embeddings with biomedical KG embeddings of Unified Medical Language System (UMLS) terms [4] extracted from MedQA questions and answer choices [14]. Our methodology introduces several key contributions. Firstly, we propose a model that integrates LLM embeddings with biomedical KG embeddings using Cross-Modal Attention (CMA) to enhance both accuracy and explainability on a biomedical reasoning task. We refer to this model as the CMA model. Our empirical results demonstrate that the CMA model outperforms the baseline of fine-tuning BioBERT alone on the MedQA dataset, highlighting the effectiveness of combining textual and KG embeddings. Additionally, we illustrate the plausible explainability of our model by visualizing the attention of the CMA mechanism.

Finding the best way to unify KGs and LLMs to combat hallucinations and enhance explainability, particularly in the biomedical domain, is an ongoing research question. We believe that our promising results may pave the way to help tackle this problem.

Related Works

Traditionally, language model predictions have been explained through attention-weight analysis. However, studies indicate that attention weights are not always reliable explanations due to the weak correlation between attention weights and input feature importance [3]. This motivates the need for more faithful explanation methods. It is demonstrated that attention on input features does not correlate well with gradient-based feature importance, counter-factual attention distributions do not change model outputs, and raw attention focuses on non-important tokens such as punctuation [13, 19]. These serve as evidence against the faithfulness and plausibility of attention as an explanation. On

the contrary, some researchers show that with certain methods, attention can correlate with gradient-based feature importance [1] and that BERT attention heads capture syntax effectively and overall attend to more than just punctuation and delimiter tokens [8]. This serves as an opportunity to explore different ways to use attention for faithful and plausible explanations, particularly on LLMs. Two promising examples make a case of utilizing attention to help explain transformer models: [6] uses the gradient of the attention weights with relevancy maps obtained through layer-wise relevance propagation [7] on bi-modal transformers to show the most important features in images for visual question and answering tasks. [20] developed a methodology that selects the most faithful attention-based interpretation by choosing the best combination of layers, matrix operations, and attention heads. They also introduce a new faithfulness metric suited for transformer models that align with ground truth rationales.

Fusing KGs and LLMs shows promising results in a variety of biomedical tasks. [28] presents QA-GNN, a method that enhances question answering by combining LLMs and KGs. It creates a joint graph representation that connects a question and answers context to entities in a retrieved knowledge subgraph, processed by an attention-based graph neural network (GNN). Deep Bidirectional Language-Knowledge Graph Pretraining (DRAGON) [27] proposes integrating KG information during the pre-training phase of LLMs with a cross-modal layer that models interactions over text and KGs, aiming to combine masked language modeling with KG link prediction. A Sophisticated Transformer Trained on Biomedical Text and Knowledge Graphs (STonKGs) [2] concatenates biomedical text with embeddings of the INDRA biomedical KG [11]. This bi-modal representation is fused in their BERT-inspired cross-encoder.

The potential of attention-based explainability and performance gains of combining LLMs with KGs inspire us to combine LLMs with KGs and utilize attention to increase model performance while increasing explainability.

Methodology

Our goal is to combine text embeddings and KG embeddings with CMA to fuse both modalities in a more explainable manner while also improving the performance of MedQA when compared to only considering the text modality.

Figure 1 illustrates an overview of our methodology. At a high level, our process involves first extracting UMLS KG triples relevant to each MedQA question and its answer choices. The triples and the MedQA text are encoded into embeddings with a graph neural network (GNN) and LLM, respectively. A single-layer cross-modal transformer that utilizes CMA (instead of self-attention) fuses both modalities. For the final prediction, a single dense layer with a softmax activation is placed on top of the classification (CLS) token's embedding that is generated by the LLM and processed through the transformer layer with CMA. This final layer is used to make predictions on MedQA. Finally, the attention probabilities of the CLS token's embedding with respect to the UMLS KG embeddings in the CMA mech-

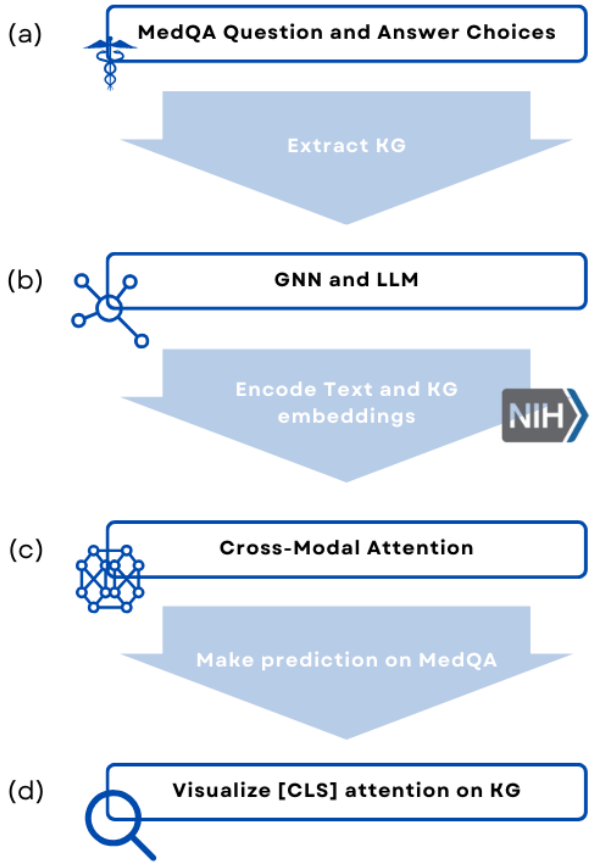

Figure 1: Overview of methodology. (a) UMLS terms and relations are extracted from MedQA text. (b) A GNN encodes these into KG embeddings and an LLM encodes the MedQA questions and answers into text embeddings. (c) A single transformer block with CMA learns the MedQA task. (d) Visualizing the CLS embedding's attention over KG embeddings gives plausible artifacts of the model's reasoning.

anism are visualized for local explainability of predictions. We will refer to this model as the CMA model.

Datasets

MedQA is a multiple choice question and answer dataset meant to gauge biomedical reasoning. MedQA has 12,723 questions that originate from medical exams [14]. The UMLS [4] KG consists of a meta-thesaurus of medical terms and their relationships represented as a KG. The UMLS KG has 297,927 UMLS terms, 98 relationships, and 1,212,586 edges. We use the UMLS KG to train our GNN and MedQA to train our CMA model.

KG Extraction

We extract KGs for each MedQA question and its answer choices. To do this we utilize SciSpacy's UMLS entity linker [21] and SemRep [15] to extract UMLS terms and relationships. We take these and form them as a set of triples that are a subset of a UMLS KG. The purpose of extracting the KGs from MedQA questions and answers is so that the CMA mechanism can explicitly fuse relevant UMLS KG embeddings with the embeddings of the MedQA text.

Text and KG Embeddings

BioBERT [16] is a BERT-like [10] LLM pre-trained over a large biomedical corpus. We use BioBERT to encode the MedQA questions and answers so that the text is represented in a way that is bidirectionally contextualized. To encode extracted KGs into KG embeddings, we inductively train a graph autoencoder on the UMLS KG for link prediction. The node features are initialized with BioBERT embeddings of their respective UMLS terms. The encoder is a 2-layer GraphSage model and the decoder is DistMult [26]. The intuition behind this is that GraphSage will enrich the original term representation with the topology of neighboring nodes [12], and DistMult [26] will further enrich the representation concerning 1-to-N and symmetric relationships within the UMLS. When producing KG embeddings for the CMA model, we only use the encoder.

Cross-modal Attention (CMA)

We fuse the text embeddings of the MedQA with the embeddings using multi-head CMA inside a transformer block [24]. This architecture is different from how StonKGs[2] utilizes CMA. We don't concatenate both modalities of embeddings and perform self-attention on the concatenated representation. Instead, our model passes the text embeddings as the attention mechanism's queries and the KG embeddings as keys and values [5, 6]. We define CMA as follows: $softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}})\mathbf{V}$ where $\mathbf{Q} \in \mathbb{R}^{h\times q\times d_h}$, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{h\times k\times d_h}$. Here, h is the number of heads, q is the number of tokens from the text, k is the number of unique UMLS terms in the extracted KG, and d_h is the embedding dimension [6]. The design choice of using a CMA allows us to fuse the two modalities in a way that conveniently allows us to visualize the explicit relationships between text and KG embeddings through analysis of the attention probabilities. Since StonKGs [2] concatenates the modalities, the direct relationships between text and KG embeddings cannot be examined explicitly.

Explanatory Visualization

Once the model is trained, the CMA mechanism is examined on MedQA predictions. [6] explores CMA between text and vision modalities to interpret visual question and answer predictions. [1, 6] averages attention across the different heads to increase the faithfulness of their attention-based explanations. These ideas are borrowed to understand how the text embeddings interact with the KG embeddings. Since the dense prediction layer is only over the final CLS token embedding, we only observe the attention probabilities produced in $softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}})$ for the CLS token embedding over the UMLS KG embeddings and we average these values across the heads. The probabilities are visualized as a $(1 \times k)$ vector with the shade of each entry corresponding to the corresponding attention probability for user readability.

Experiment

Datasets

To finetune the baseline LLM and train our CMA model, we use the MedQA dataset as described above.

Training details

The baseline model, BioBERT [16], is fine-tuned on MedQA using the Huggingface transformers library. [25].

To train the CMA model we set $h = 8$ and $d_h = 768$ in the CMA mechanism. We minimize the cross-entropy loss with the AdamW optimizer [17] paired with a linear scheduler with 100 warm-up steps and the learning rate $lr = 5e-5$. For the CMA model, we use pre-trained embeddings of BioBERT obtained using the Huggingface transformers library [25]. At each step of the training process, the CMA model processes two main inputs: first, the BioBERT embeddings derived from a MedQA dataset question and its answer choices, and second, the KG that has been extracted from the same MedQA dataset entry that has been encoded into KG embeddings by the GNN. It is also important to note that in training, no additional learning takes place for the GNN or BioBERT, they are only used to generate embeddings for the CMA model to input.

Model	val_acc	test_acc
BioBERT	0.291	0.279
CMA	**0.296**	**0.305**

Table 1: Comparison of model accuracy on the MedQA dataset.

Ablation study

We conduct an ablation study to understand the CMA model's and BioBERT's difference in performance and generalization on the MedQA dataset. We use validation and test accuracies as our evaluation criteria. Table 1 outlines the experiment results obtained from this study.

Baselines. The baseline we use is BioBERT fine-tuned on MedQA. We compare the results of the baseline with the CMA model trained on MedQA.

Results. BioBERT finetuned on MedQA exhibits validation and test accuracies of 0.291 and 0.279, respectively. Interestingly, the CMA model resulted in an enhancement in performance, achieving validation and test accuracies of 0.296 and 0.305, respectively. Moreover, the combination of BioBERT with CMA not only elevated the accuracy but also enhanced the model's generalization capability. This is evidenced by the fact that the CMA model's performance on the test dataset surpassed its performance on the validation dataset. BioBERT's test accuracy is worse than its validation accuracy. These results indicate a promising direction for improving both the accuracy and generalizability of LLMs through the strategic use of CMA.

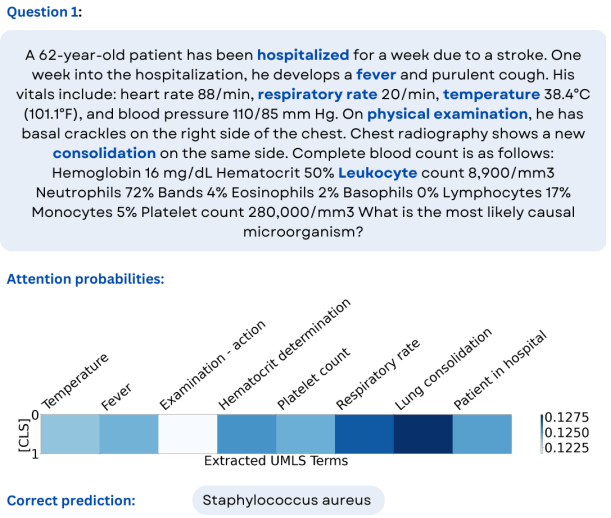

Figure 2: Example of a visualized explanation where the CMA model makes a correct prediction on an unseen MedQA question. The blue highlighted words are where UMLS terms are extracted from the questions and answers. The plots illustrate the attention probabilities of the CLS token over the extracted UMLS terms. A darker color indicates a higher attention probability.

Explanations

As opposed to self-attention-based interpretations [1, 20], which either only show how tokens attend to other tokens [8] or does not explicitly show the relationship between the text and KG modalities [2], we present a different viewpoint to interpret the predictions made by our CMA model.

Instead, we can explicitly examine the relationships between the fused BioBERT and KG embeddings by observing the CMA's attention probabilities. Our method of explanation is inspired by [1] averaging attention across the attention heads and [6] using CMA and visualizing attention for visual question and answer models.

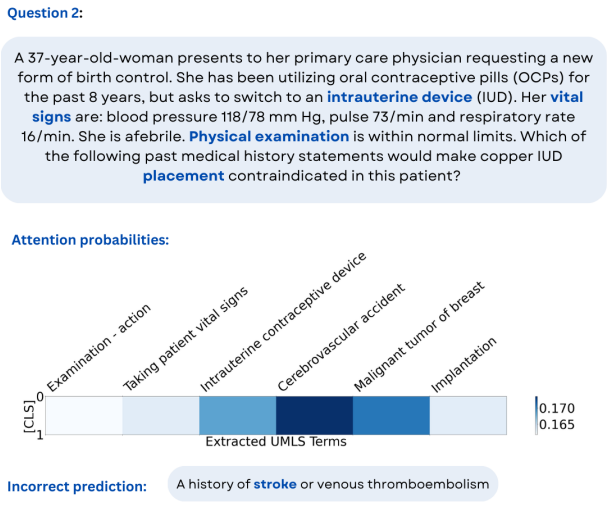

Figure 3: Example of a visualized explanation where the CMA model makes an incorrect prediction on an unseen MedQA question. The blue highlighted words are where UMLS terms are extracted from the questions and answers. The plots illustrate the attention probabilities of the CLS token over the extracted UMLS terms. A darker color indicates a higher attention probability.

The crux of our explanation methodology lies in visualizing the attention probabilities of the model's CLS token embedding over the different KG embeddings. We chose the CLS token because the CMA model's dense prediction layer inputs the CLS token embeddings. This visualization serves as a window into the model's reasoning, providing insights into how it associates biomedical concepts encoded in the KG with the text of the question and answer choices. Figures 2 and 3 exemplify this, showcasing the model's attention patterns for both correct and incorrect predictions.

When the model makes a correct prediction, such as identifying 'Staphylococcus Aureus' (Figure 2), a microorganism known to cause pneumonia, the attention visualization reveals a strong focus on clinically relevant UMLS terms extracted from the question and answer choices, in this case, 'Lung consolidation'. This example is significant since pneumonia is a type of lung consolidation [9, 23].

Conversely, when the model makes incorrect predictions, visualizing the CMA offers critical insights into the model's potential areas of confusion. For example, when the model incorrectly predicts 'A history of stroke or venous thromboembolism' (Figure 3), the disproportionate attention to 'Cerebrovascular accident' (the medical term for stroke) indicates a misunderstanding or an overemphasis on certain terms. This analysis is invaluable for identifying and addressing the model's biases or knowledge gaps.

These findings illustrate the potential of CMA not only to enhance model performance but also to offer plausible explanations for the model's predictions. The benefit of this type of visualization is that humans can easily understand which UMLS terms the CLS token attends to the most and

trace back where from the original text the UMLS terms are extracted. This joint benefit is particularly promising in the biomedical domain, where both accuracy and explainability are crucial.

Conclusion

In conclusion, our research demonstrates the significant potential of integrating LLMs with KGs to enhance both the performance and the explainability of models in the biomedical domain. Other works in this field demonstrate increased performance but do not address the problem of attention-based explainability. We show that by using CMA to fuse an LLM and KG, not only do we achieve improved accuracy on a biomedical reasoning task but also demonstrate plausible attention-based explanations for local predictions.

Looking ahead, we aim to further refine our approach by comparing it to different baseline LLMs, addressing the faithfulness of our explanations, optimizing the graph extraction process, and expanding our model's applicability to a broader range of tasks. Additionally, a comprehensive analysis of each component's contributions through ablation studies will deepen our understanding of the model's mechanics and inform future enhancements.

Our work contributes to the growing body of research advocating for the integration of LLMs and KGs. As we continue to explore this promising avenue, we remain committed to unlocking new potentials in bio-medicine, aiming for a future where bio-medicine and artificial intelligence can converge more transparently and effectively.

References

[1] Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

[2] Balabin, H.; Hoyt, C. T.; Birkenbihl, C.; Gyori, B. M.; Bachman, J.; Kodamullil, A. T.; Plöger, P. G.; Hofmann-Apitius, M.; and Domingo-Fernández, D. 2022. STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs. *Bioinformatics*, 38(6): 1648–1656.

[3] Bibal, A.; Cardon, R.; Alfter, D.; Wilkens, R.; Wang, X.; François, T.; and Watrin, P. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3889–3900.

[4] Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1): D267–D270.

[5] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

[6] Chefer, H.; Gur, S.; and Wolf, L. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 397–406.

[7] Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782–791.

[8] Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

[9] Clark, S.; and Hicks, M. 2023. Staphylococcal pneumonia.[Updated 2022 Aug 8]. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*.

[10] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[11] Gyori, B. M.; Bachman, J. A.; Subramanian, K.; Muhlich, J. L.; Galescu, L.; and Sorger, P. K. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology*, 13(11): 954.

[12] Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

[13] Jain, S.; and Wallace, B. C. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

[14] Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

[15] Kilicoglu, H.; Rosemblat, G.; Fiszman, M.; and Shin, D. 2020. Broad-coverage biomedical relation extraction with SemRep. *BMC bioinformatics*, 21: 1–28.

[16] Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

[17] Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[18] Meskó, B.; and Topol, E. J. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1): 120.

[19] Mohankumar, A. K.; Nema, P.; Narasimhan, S.; Khapra, M. M.; Srinivasan, B. V.; and Ravindran, B. 2020. Towards transparent and explainable attention models. *arXiv preprint arXiv:2004.14243*.

[20] Mylonas, N.; Mollas, I.; and Tsoumakas, G. 2024. An attention matrix for every decision: faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification. *Data Mining and Knowledge Discovery*, 38(1): 128–153.

[21] Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319–327.

Florence, Italy: Association for Computational Linguistics.

[22] Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *arXiv preprint arXiv:2306.08302*.

[23] Team, T. H. E. 2023. Lung consolidation: Treatment, vs pleural effusion, and more.

[24] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[25] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[26] Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

[27] Yasunaga, M.; Bosselut, A.; Ren, H.; Zhang, X.; Manning, C. D.; Liang, P. S.; and Leskovec, J. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35: 37309–37323.

[28] Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.