REVELIO: INTERPRETABLE LONG-FORM QUESTION ANSWERING

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Fabian Vincenzi & Davide Freddi
Department of Computer Science and Engineering (DISI)
University of Bologna, Via dell'Università 50, I-47522 Cesena, Italy
{gianluca.moro, l.ragazzi, lorenzo.valgimigli}@unibo.it
{fabian.vincenzi, davide.freddi3}@studio.unibo.it

ABSTRACT

The black-box architecture of pretrained language models (PLMs) hinders the interpretability of lengthy responses in long-form question answering (LFQA). Prior studies use knowledge graphs (KGs) to enhance output transparency, but mostly focus on non-generative or short-form QA. We present REVELIO, a new layer that maps PLM's inner working onto a KG walk. Tests on two LFQA datasets show that REVELIO supports PLM-generated answers with reasoning paths presented as rationales while retaining performance and time akin to their vanilla counterparts.

1 INTRODUCTION

Closed-book long-form question answering (LFQA) asks pretrained language models (PLMs) to generate long responses using only the knowledge stored in their parameters. While avoiding open-book passage retrieval accelerates training and inference, the black-box nature of PLMs hampers answer interpretation, limiting real-world applicability. To improve human understanding (Commission, 2020), prior studies use knowledge graphs (KGs) to provide logical reasoning (Yasunaga et al., 2021; Zhang et al., 2022b). However, they focus on non-generative multiple-choice QA (MCQA), neglecting the practicality of LFQA (see Appendix A). We introduce REVELIO, a plugin layer that enables PLMs to craft answers by aligning parametric knowledge with an external KG, offering reasoning paths as support for decision interpretation. Our study delves into two closed-book LFQA datasets by testing two million-scale PLMs. Quantitative and qualitative results show that REVELIO provides reasoning pathways as the rationale behind a given answer while marginally improving performance. Our code is open at https://disi-unibo-nlp.github.io/projects/revelio/.

2 METHOD

We introduce REVELIO, our proposed plugin layer to allow PLMs to communicate with an external KG to provide answer-related reasoning paths to enhance interpretability (Figure 1).

Preprocessing. Given a question x, we use RAKE (Rose et al., 2010) to extract a set of keywords $\mathcal{K} = \{k_1, \dots, k_{|\mathcal{K}|}\}$. We create $|\mathcal{K}|$ depth-d subgraphs $\mathcal{G} = \{g_1, \dots, g_{|\mathcal{K}|} \mid g_i = \langle \mathcal{N}_r, \mathcal{E}, \mathcal{N}_e \rangle\}$, setting k_i as the root node of g_i, where d is a hyperparameter and $\langle \mathcal{N}_r, \mathcal{E}, \mathcal{N}_e \rangle$ are relational triplets with $\mathcal{N}_r$, $\mathcal{N}_e$, and $\mathcal{E}$ representing the root node, end node, and edges, respectively. For each k, we perform an exact match on CONCEPTNET (Speer et al., 2016) to retrieve the corresponding depth-d subgraph starting from k. Following Hu et al. (2022), we augment x by preceding each k with new tokens <rel_tok> and <node_tok> to allow PLMs to interact with $\mathcal{E}$ and $\mathcal{N}$, respectively. We then define the REVELIO graph walk $\mathcal{W} = \{w_1, \dots, w_{|\mathcal{K}|}\}$, where w_i is the path on g_i starting from k_i. Iteratively, REVELIO adds a triplet (e.g., ⟨*water*, *is-a*, *liquid*⟩) to each w_i, symbolizing a step in the walk. For simplicity, we define the end node of the last triplet as *Current Node*, representing the position of the model in the graph. Appendix C reports in-depth elucidation and ablation studies.

Execution. For each w_i, REVELIO aims to select the next most salient node and inject the new information from $\mathcal{G}$ into $\mathcal{H}_x$—the last hidden state of x. To streamline, we explain the process for a single k, but the same mechanism is run in parallel for each $k \in \mathcal{K}$. The input of REVELIO is $(\mathcal{H}_x, \mathcal{G}, \mathcal{W})$.

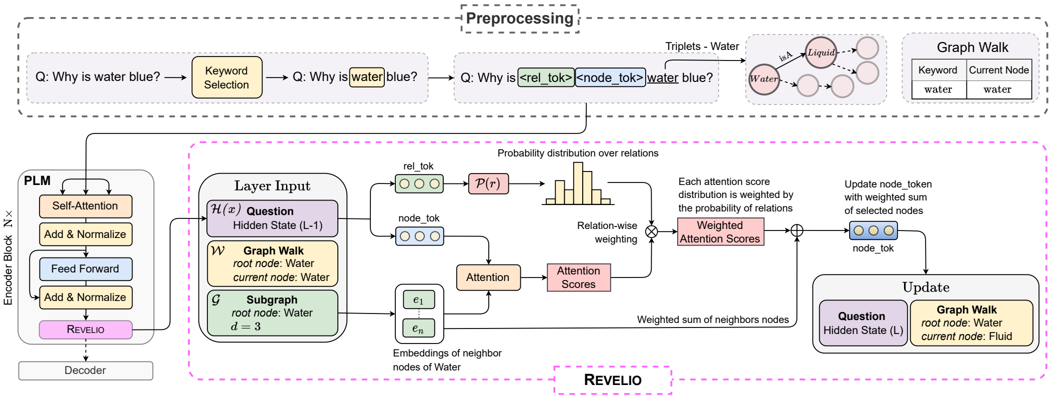


Figure 1: The overview of REVELIO, our proposed approach.

Model (size)	ELI5 (2019) R1	R2	RL	$\mathcal{R}$	BeS	BaS	AQUAMUSE (2020) R1	R2	RL	$\mathcal{R}$	BeS	BaS
T5 (base)	22.93	4.11	14.02	13.61	2.91	-2.48	30.09	7.98	19.67	19.09	**11.61**	**-2.29**
w/ REVELIO	**23.01**	**4.26**	**14.17**	**13.73**	**3.10**	**-2.47**	**30.14**	**8.18**	**20.42**	**19.42**	11.05	-2.43
T5 (large)	23.35	4.21	**14.25**	13.85	**5.35**	-2.31	**30.87**	8.95	**20.74**	**20.03**	**17.52**	-2.17
w/ REVELIO	**23.57**	**4.26**	14.16	**13.91**	5.06	**-2.30**	30.62	**9.17**	20.45	19.93	15.58	**-2.11**
BART (base)	23.21	3.33	13.16	13.15	-15.32	-5.40	24.11	4.72	17.14	15.23	-11.72	-5.41
w/ REVELIO	**23.41**	**3.74**	**13.27**	**13.39**	**-14.22**	**-5.27**	**24.67**	**4.78**	**17.32**	**15.49**	**-10.98**	**-5.21**
BART (large)	24.23	4.17	14.71	14.27	-7.54	-3.06	**26.67**	5.98	**19.52**	**17.26**	-5.28	**-3.41**
w/ REVELIO	**24.62**	**4.23**	**15.12**	**14.56**	**-7.22**	**-2.58**	26.53	**6.12**	19.41	17.23	**-5.11**	-3.59

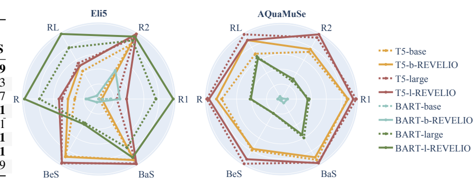


Figure 2: Results on the benchmarked datasets. The best intra-model score in the table is in bold.

First, we extract e_r, e_n from $\mathcal{H}_x$, i.e., the <rel_tok> and <node_tok> embeddings, respectively. We then linearly project e_r and apply a softmax operation to yield a probability distribution $\mathcal{P}(r)$ over all possible edges (24 types in CONCEPTNET, e.g., *is-a*, *is-composed-of*). Meanwhile, we perform an attention operation to compare e_n with all neighbor nodes n of the *Current Node* of w, producing a score s_e for each pair. The scores are then weighted by $\mathcal{P}(r)$, considering the relation that links the *Current Node* and n. The node with the highest score is added to $\mathcal{W}$. Then, all node embeddings, weighted by their scores, are summed to e_n to inject KG information into the PLM.

3 EXPERIMENTS

Setup. We train and evaluate the two most popular million-scale PLMs, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), with and without REVELIO on two closed-book public LFQA datasets: ELI5 (Fan et al., 2019) and AQUAMUSE (Kulkarni et al., 2020). For all benchmarks, we automatically evaluate models by reporting % of the F1 scores of ROUGE-{1,2,L} (Lin, 2004) and $\mathcal{R}$ (Moro et al., 2023b) for syntactic matching, BERTScore (BeS) (Zhang et al., 2020) for semantic assessment, and BARTScore-$\mathcal{F}$ (BaS) (Yuan et al., 2021) to judge factuality. We perform human analysis using a direct comparison strategy that has been proven to be more reliable and less labor-intensive than rating scales (Huang et al., 2023; Moro et al., 2023d). More details are in Appendix B.

Results. Figure 2 shows the overall results of REVELIO. In detail, the adoption of REVELIO overall slightly improves model performance across datasets and metrics in two different LFQA tasks. This finding indicates the benefit of interacting with KGs to extract contextualized information. Human annotation (with an agreement of 78%) shows that 85% of the time REVELIO's answers are comparable or better than those of T5 (see Appendix B for details). Ablation studies are given in Appendix C. Finally, graphical examples of reasoning paths are provided in Appendix D.

4 CONCLUSION

We present REVELIO, a flexible layer to enhance the output of current PLMs with interpretable reasoning paths. Experiments on two closed-book LFQA datasets show that models equipped with REVELIO generate better answers than their vanilla counterparts, paving the way for novel promising KG interaction methods for LFQA. Limitations and future directions are discussed in Appendix E.

Acknowledgements

This research is partially supported by (i) the Complementary National Plan PNC-I.1, "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, DARE—DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (ii) the PNRR, M4C2, FAIR—Future Artificial Intelligence Research, Spoke 8 "Pervasive AI," funded by the European Commission under the NextGeneration EU program. The authors thank the Maggioli Group[1] for granting the Ph.D. scholarship to Luca Ragazzi and Lorenzo Valgimigli.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

European Commission. On artificial intelligence—a european approach to excellence and trust, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL `https://doi.org/10.18653/v1/n19-1423`.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 3558–3567. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1346. URL `https://doi.org/10.18653/v1/p19-1346`.

Giacomo Frisoni and Gianluca Moro. Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In Slimane Hammoudi, Christoph Quix, and Jorge Bernardino (eds.), *Data Management Technologies and Applications - 9th International Conference, DATA 2020, Virtual Event, July 7-9, 2020, Revised Selected Papers*, volume 1446 of *Communications in Computer and Information Science*, pp. 293–318. Springer, 2020. doi: 10.1007/978-3-030-83014-4_14. URL `https://doi.org/10.1007/978-3-030-83014-4_14`.

Xin Guan, Biwei Cao, Qingqing Gao, Zheng Yin, et al. CORN: co-reasoning network for commonsense question answering. In *COLING*, pp. 1677–1686. ICCL, 2022.

Chuzhan Hao, Minghui Xie, and Peng Zhang. Acenet: Attention guided commonsense reasoning on hybrid knowledge graph. In *EMNLP*, pp. 8461–8471. ACL, 2022.

Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, et al. Empowering language models with knowledge graph reasoning for open-domain question answering. In *EMNLP*, pp. 9562–9581. ACL, 2022.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. SWING: Balancing coverage and faithfulness for dialogue summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 512–525, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.37. URL `https://aclanthology.org/2023.findings-eacl.37`.

[1] `https://www.maggioli.com/who-we-are/company-profile`

Jinhao Jiang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. $great truths are always simple: $ A rather simple knowledge encoder for enhancing the commonsense reasoning capacity of pre-trained models. In *NAACL-HLT (Findings)*, pp. 1730–1741. Association for Computational Linguistics, 2022.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *CoRR*, abs/2010.12694, 2020. URL https://arxiv.org/abs/2010.12694.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL 2020, Online, July 5-10, 2020*, pp. 7871–7880. ACL, 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://doi.org/10.18653/v1/2020.acl-main.703.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. ACL. URL https://aclanthology.org/W04-1013.

Gianluca Moro and Gabriele Monti. W-grid: A scalable and efficient self-organizing infrastructure for multi-dimensional data management, querying and routing in wireless data-centric sensor networks. *J. Netw. Comput. Appl.*, 35(4):1218–1234, 2012. doi: 10.1016/J.JNCA.2011.05.002. URL https://doi.org/10.1016/j.jnca.2011.05.002.

Gianluca Moro and Luca Ragazzi. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 11085–11093. AAAI Press, 2022. doi: 10.1609/AAAI.V36I10.21357. URL https://doi.org/10.1609/aaai.v36i10.21357.

Gianluca Moro and Luca Ragazzi. Align-then-abstract representation learning for low-resource summarization. *Neurocomputing*, 548:126356, 2023. doi: 10.1016/J.NEUCOM.2023.126356. URL https://doi.org/10.1016/j.neucom.2023.126356.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 180–189. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.15. URL https://doi.org/10.18653/v1/2022.acl-long.15.

Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and Paolo Italiani. Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law*, pp. 1–29, 2023a. doi: 10.1007/s10506-023-09373-8. URL https://doi.org/10.1007/s10506-023-09373-8.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 14417–14425. AAAI Press, 2023b. doi: 10.1609/AAAI.V37I12.26686. URL https://doi.org/10.1609/aaai.v37i12.26686.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. Graph-based abstractive summarization of extracted essential knowledge for low-resource scenarios. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Radulescu (eds.), *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 1747–1754. IOS Press, 2023c. doi: 10.3233/FAIA230460. URL https://doi.org/10.3233/FAIA230460.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Giacomo Frisoni, Claudio Sartori, and Gustavo Marfia. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*, 23(7):3542, 2023d. doi: 10.3390/S23073542. URL `https://doi.org/10.3390/s23073542`.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Lorenzo Molfetta. Retrieve-and-rank end-to-end summarization of biomedical studies. In Oscar Pedreira and Vladimir Estivill-Castro (eds.), *Similarity Search and Applications - 16th International Conference, SISAP 2023, A Coruña, Spain, October 9-11, 2023, Proceedings*, volume 14289 of *Lecture Notes in Computer Science*, pp. 64–78. Springer, 2023e. doi: 10.1007/978-3-031-46994-7_6. URL `https://doi.org/10.1007/978-3-031-46994-7_6`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pp. 1–20, 2010.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975, 2016. URL `http://arxiv.org/abs/1612.03975`.

Dhaval Taunk, Lakshya Khanna, Siri Venkata Pavan Kumar Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi. Grapeqa: Graph augmentation and pruning to enhance question-answering. In *WWW (Companion Volume)*, pp. 1138–1144. ACM, 2023.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, et al. QA-GNN: reasoning with language models and knowledge graphs for question answering. In *NAACL-HLT*, pp. 535–546. ACL, 2021.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In *NeurIPS*, 2022.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, pp. 27263–27277, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html`.

Miao Zhang, Rufeng Dai, Ming Dong, and Tingting He. DRLK: dynamic hierarchical reasoning with language model and knowledge graph for question answering. In *EMNLP*, pp. 5123–5133. ACL, 2022a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, et al. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, et al. Greaselm: Graph reasoning enhanced language models for question answering. *CoRR*, abs/2201.08860, 2022b.

A RELATED WORK

Currently, there is strong interest in explaining the reasoning behind the output of PLMs to enhance interpretability (Frisoni & Moro, 2020). To this end, the synergy between KGs and PLMs has been explored in the field of QA. QA-GNN (Yasunaga et al., 2021) extracts subgraphs for each possible answer, and a multi-layer perceptron assigns the probability of being the correct answer. GREASE-LM (Zhang et al., 2022b) extends the work by creating connections between KG and PLM. Other solutions, such as CORN (Guan et al., 2022) and ACENET (Hao et al., 2022), improve graph processing with PLM information and vice versa using different attention mechanisms. SAFE (Jiang et al., 2022) simplifies the process by considering only relations, not nodes, proving that the relationships

Table 1: Related works about the integration of QA and KG.

Model	Task	KG*	Plug†	Gen‡	PLM (Size)
QA-GNN (Yasunaga et al., 2021)	MCQA	ConceptNet, UMLS, DrugBan	✗	✗	RoBERTa (large), AristoRoBERTa
GreaseLM (Zhang et al., 2022b)	MCQA	ConceptNet, Desease DB	✗	✗	RoBERTa (large), SapBERT
CORN (Guan et al., 2022)	MCQA	ConceptNet	✗	✗	RoBERTa (large)
ACENet (Hao et al., 2022)	MCQA	ConceptNet	✗	✗	RoBERTa (large), AristoRoBERTa, SapBERT
DRKL (Zhang et al., 2022a)	MCQA	ConceptNet, Desease DB	✗	✗	RoBERTa (large), SapBERT
Dragon (Yasunaga et al., 2022)	MCQA	Freebase, Wikidata, ConceptNet	✗	✓	RoBERTa (large)
GrapeQA (Taunk et al., 2023)	MCQA	ConceptNet	✗	✗	RoBERTa (large), SapBERT
SAFE (Jiang et al., 2022)	MCQA	ConceptNet	✓	✗	RoBERTa (large), AristoRoBERTa, BERT(large)
OreoLM (Hu et al., 2022)	MCQA	ConceptNet	✓	✓	RoBERTa, T5(base/large)
Revelio *(Ours)*	LFQA	ConceptNet	✓	✓	T5 (base/large), Bart (base/large),

* The external knowledge graphs used.
† ✓ = the solution can be plugged into different PLMs with minimal effort and the paper contains experiments on that; ✗ = otherwise.
‡ ✓ = the output is generated (i.e., abstractive); ✗ = otherwise.

Table 2: The number of trainable parameters of PLMs.

Model	Parameters	URL
Bart-base	140M	`https://huggingface.co/facebook/bart-base`
Bart-large	400M	`https://huggingface.co/facebook/bart-large`
T5-base	220M	`https://huggingface.co/t5-base`
T5-large	770M	`https://huggingface.co/t5-large`
Revelio	2M	-

are sufficient for this task. The paradigm change is given by Dragon (Yasunaga et al., 2022), a generative model pretrained by coupling masked language modeling and link prediction, outperforming all previous models in the QA task. So far, all models work by generating a probability over the possible answers. Therefore, the applicability is narrowed to answer selection, i.e., MCQA, leaving generative QA unexplored. OreoLM (Hu et al., 2022) tries to mitigate this problem using a generative model to produce a short answer. However, all these existing solutions focus on MCQA, with the aim of guessing the most probable answer given a set of alternatives or generating an answer composed of a few words. In this work, we address the generation of long and complex answers, leveraging a KG to improve model performance and interpret the reasoning. Table 1 compares our contribution with previous works.

B Experimental Details

Models. Bart is a transformer with quadratic memory and time complexity in input size characterized by a denoising pretraining objective. T5 is a quadratic transformer with a text-to-text objective. Table 2 lists the number of parameters and the URL of the model checkpoints.

Datasets. Eli5 comprises question–answer pairs extracted from the Reddit forum "Explain Like I'm Five." AQuaMuSe collects query-based summaries of topic-related documents. In our experiments, given the input question, we directly use the summary as the answer without providing the source documents to the models. Table 3 provides the dataset statistics.[2]

[2]All datasets are publicly available in Hugging Face: `https://huggingface.co/datasets/eli5` and `https://huggingface.co/datasets/aquamuse`.

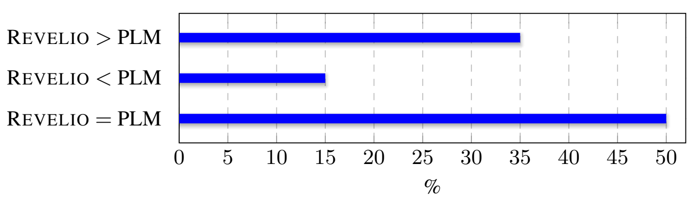
Table 3: Statistics of the datasets used as testbeds. All values are averaged except "# Instances."

Dataset	Domain	# Train	# Dev	# Test	Source # Words	Target # Words
ELI5 (2019)	Commonsense	5000	100	1000	42.2	130.6
AQUAMUSE (2020)	Commonsense	4555	440	524	15.5	105.9

Table 4: The settings of the evaluation metrics.

Metric	Description	Bound*	Hyperparameters
ROUGE-1/2/L (Lin, 2004)	Lexical overlaps of unigrams (R-1), bigrams (R-2), and longest common subsequence (R-L).	$[0, 1], \uparrow$	rouge_types=["rouge1","rouge2","rougeL"], use_stemmer=True
$\mathcal{R}$ (Moro et al., 2023b)	Aggregated ROUGE score penalizing results with discrepant R-1, R-2, R-L.	$[0, 1], \uparrow$	/
BERTScore (Zhang et al., 2020)	IDF-weighted n-gram alignment through contextualized embeddings from BERT (Devlin et al., 2019).	$[-1, 1], \uparrow$	model_type="microsoft/deberta-large-mnli", rescale_with_baseline=True, batch_size=32
BARTScore-$\mathcal{F}$ (Yuan et al., 2021)	Estimation of BART (Lewis et al., 2020) of how predictions and references are mutual paraphrases.	$[-\infty, 0], \uparrow$	checkpoint="facebook/bart-large-cnn", batch_size=4

* $\uparrow$ = the higher, the better.



Comparison	Score
REVELIO $>$ PLM	**35%**
REVELIO $<$ PLM	15%
REVELIO $=$ PLM	50%

Figure 3: Human evaluation results on 40 random samples of ELI5.

Metrics. Table 4 shows technical details on the automatic evaluation metrics used. Regarding human annotation, we provide three English-proficient QA researchers with 40 random test set questions from ELI5 and their corresponding gold and machine-generated answers from T5-base. We ask the annotators to state which answer is the best in terms of correctness and informativeness w.r.t. the given question. A tie is declared if the judge perceives that the two answers are of comparable quality. We randomize the order of instances to guard the rating against being gamed. We collect the responses and aggregate them using majority voting. The results are depicted in Figure 3.

Implementation Details. We fine-tuned the models using PyTorch and the HuggingFace library, setting the seed to 42 to guarantee reproducibility. All PLMs are trained for 5 epochs with a learning rate of $3e^{-5}$, using mixed precision and gradient checkpointing to preserve memory. We selected the checkpoint that performed the best on the validation set at the end of each training epoch. Our REVELIO layer uses a different optimizer whose learning rate is $1e^{-6}$ (see Appendix C for particulars). At inference time, we use the beam search decoding with 3 beams, n-gram repetition blocks for n>3, and an output length in $[50, 140]$ tokens.

Hardware Configuration. We used a workstation with 4 Nvidia GeForce RTX3090 GPU of 24 GB memory, 64 GB of VRAM, and an Intel® Core™ i9-10900X1080 CPU @ 3.70GHz processor.

C ABLATION STUDIES

To optimize hyperparameter selection, we conducted an extensive series of experiments using OPTUNA,[3] a sophisticated open-source tool designed for hyperparameter optimization that inherently supports parallelism. We configure OPTUNA to seamlessly integrate with SLURM, our resource management system. This configuration allowed us to run parallel experiments efficiently on four

[3] https://optuna.org/

Table 5: ROUGE scores with different hyperparameters. *Left*: a comparison of the number of REVELIO layers (mean of up to 10 different settings). *Right*: different keyword extraction methods.

# Layers	R1	R2	RL
1	20.65	2.81	10.32
2	**21.83**	3.45	12.71
3	21.81	**3.59**	**12.78**
4	20.93	3.06	12.21

Method	R1	R2	RL
RAKE	21.64	**3.71**	**13.09**
JAKE	21.23	3.32	12.89
KEYBERT	20.79	3.02	11.97
YAKE	**21.67**	3.44	12.76

GPUs, while collecting all the data on a MySQL server located on a separate node for optimized data handling. Table 5 reports the results of the ablation studies, performed with T5-base on ELI5.

Learning Rate Optimization. We started by identifying the optimal learning rate for training the PLM with REVELIO. We differentiated between two distinct learning rates: $lr_{\mathcal{R}}$ for custom layers and $lr_{\mathcal{P}}$ for the standard parameters of the model. Initially, we considered employing separate optimizers for each set of parameters; however, empirical evidence suggested that it did not provide a significant advantage. Consequently, we adopt a unified optimization approach. Through rigorous testing, we determined that the most effective training occurred with $lr_{\mathcal{R}} = 1e^{-6}$ and $lr_{\mathcal{P}} = 3e^{-5}$.

REVELIO Layers. We experimented using REVELIO in different layers. We let OPTUNA decide how many layers to use from 1 to 4, and which layers to alter. Our findings revealed that the incorporation of 2 or 3 REVELIO layers yielded comparable effective results. Given this equivalence in performance, we opt for 2 layers to minimize computational complexity and reduce the depth of the graph $\mathcal{G}$ by 1 layer, thus streamlining the process. Subsequently, we conducted a more focused experiment, fixing the layer count at 2 while using OPTUNA to identify the most impactful layer positions. The most effective configuration emerged as a combination of the 3rd and 7th layers, striking a balance between the early and later stages of information processing within the PLM.

Keyword Extraction Algorithm. Another key component investigated in the early phase was the keyword extractor. We tested RAKE, JAKE, YAKE, and KEYBERT, which are state-of-the-art methods for keyword extraction. We up-limit the number of keywords to 5 via hyperparameter search to mitigate possible noise. We observed that RAKE produced the most favorable results (see Table 5). Regarding the exact match over CONCEPTNET, if the exact match of k fails, we drop k from $\mathcal{K}$ (a scenario that occurs on average 1% of the time).

D OUTPUT ANALYSIS

Figure 4 shows graphical examples of reasoning paths. Table 6 reports some erroneous predictions.

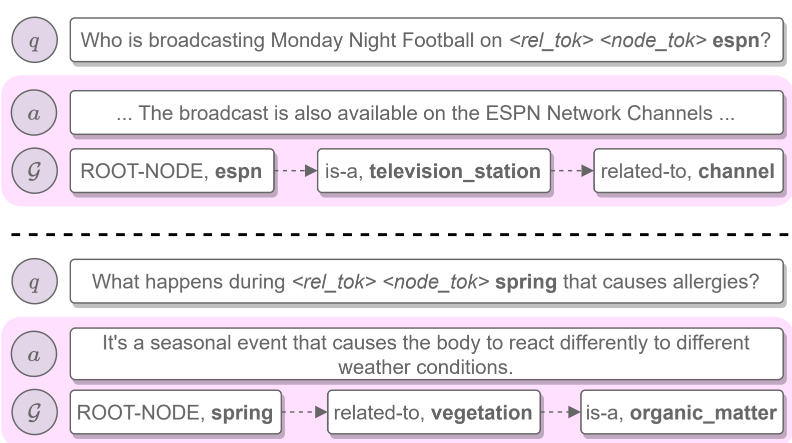


Figure 4: Graphical examples of reasoning paths from ELI5.

Table 6: Erroneous output examples from ELI5.

#1 Question
How come we can not feel the heat of the inner earth if we are only 18 miles above it?
#1 Gold Answer
For the same reason we don't burn to death when a steel forge is fired up from a mile away. The asthenosphere's temperature is between 300 and 500 degrees Celsius. The earth's crust, from the surface to the asthenosphere, is 30 to 50 miles thick. What's more, the earth's crust has an extremely low thermal conductivity. No material exists that has a remotely sufficient specific heat capacity to transfer 500C of heat through 18 miles of (highly insulating) earth.
#1 Generated Answer
The heat of the inner earth is absorbed by the air. It's not like we can't feel it, but it's just that we're only 18 miles above it. If you're in a high altitude, you can feel it.

#2 Question
God is a woman ariana grande writing credits?
#2 Gold Answer
"God Is a Woman" (stylized "God is a woman") is a song by American singer Ariana Grande. It was released on July 13, 2018, as the second single from Grande's fourth studio album Sweetener (2018). The song was written by Grande, Max Martin, Savan Kotecha, Rickard Göransson and its producer Ilya.
#2 Generated Answer
God is a woman, a writer, and an actress. She is best known for her role in the television series The Greatest Showman, which premiered on November 5, 2018. She is also known for playing the role of a female character in the TV series The Big Bang Theory.

E LIMITATIONS AND FUTURE DIRECTIONS

Future work should explore the following limitations.

Computational Intensity. Our solution requires the creation of a graph for each question, which is not demanding at experimental time (i.e., for training and testing in existing datasets) because graph creation is performed offline as a preprocessing phase. In contrast, in a real-world application characterized by a flow of new user questions, there will be a waiting time due to this procedure.

Knowledge Base Dependence. Our method requires a knowledge base to work, which could be challenging to find or create for a diverse range of cases, such as biomedical (Moro et al., 2022; 2023e) and legal applications (Moro & Ragazzi, 2022; Moro et al., 2023a).

Reasoning Paths Evaluation. There are no automatic evaluation metrics that can capture the usefulness of the reasoning paths provided. Even human analysis is not as simple as it may seem because of the lack of a rigorous standard to follow.

Additional Tasks. As introduced in communication networks (Moro & Monti, 2012), tracking and propagating knowledge refinements between graph nodes can beneficial for creating interpretable pathways to support prediction in other real-world generative tasks such as text summarization (Moro & Ragazzi, 2023; Moro et al., 2023c).