Conversational Equations: A Conversational Question-Answering Dataset Grounded in Scientific Equations

Anirudh Sundar, Shaleen Parikh, Larry Heck
AI Virtual Assistant Laboratory, Georgia Institute of Technology
{asundar34, shaleen, larryheck}@gatech.edu

Abstract

An emerging area of research in situated conversational AI is the creation of a Virtual Research Assistant (VRA). The VRA is a conversational agent that supports and amplifies human research. Among other challenges, the VRA must be capable of contextual dialogue grounded in scientific papers. An important element of conversational scientific papers is interpreting document-grounded equations to support an open dialogue question-answering interaction with the human researcher. This work introduces CONVERSATIONAL EQUATIONS (cEQNS), a dataset of multi-turn conversational question-answer pairs grounded in equations and their associated references from scientific documents available on arXiv.

1 Introduction

An ongoing challenge in AI research is the development of conversational assistants that effectively engage in dialogue using structured knowledge (Sundar and Heck, 2022), particularly in handling scientific literature via a Virtual Research Assistant (VRA). This task is critical due to the ever-increasing volume of scientific papers and the complexity of their multimodal content, including text, images of models and processes, tables and charts for data comparison, and mathematical equations.

In particular, equations are crucial for grasping mathematical concepts in scientific texts but can be challenging to interpret, especially for beginners or with new formulations. They frequently rely on notation introduced elsewhere in the document, and require readers to review the entire text.

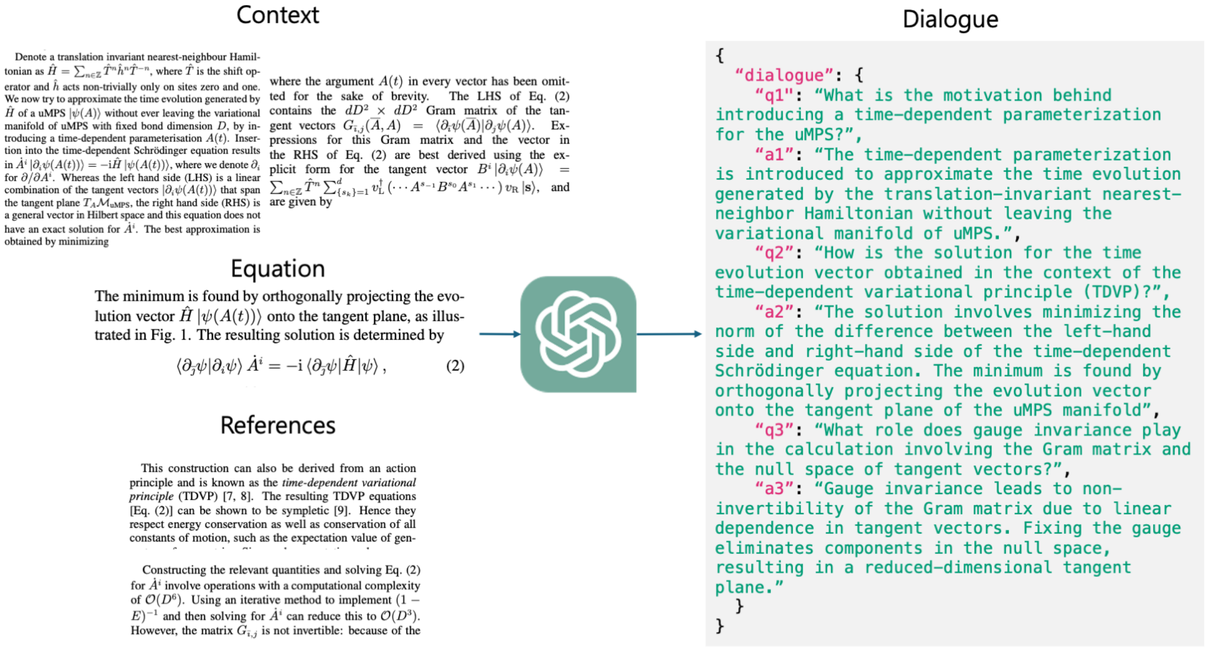
Motivated by these challenges, the task of modeling mathematical equations and natural language text has become a topic of active research. Prior work has focused on retrieving equations, generating natural language text conditioned on equations, grounding equations in descriptions, and solving math word problems (Chiang and Chen, 2019; Wang et al., 2021a; Peng et al., 2021). However, a primary challenge in understanding mathematical equations is to build a VRA capable of answering questions in a conversational context, the construction of which necessitates a dataset of conversations situated in document-grounded mathematical equations which does not yet exist.

To address this issue, we introduce CONVERSATIONAL EQUATIONS (cEQNS), a dataset featuring conversational QA pairs linked to mathematical equations and references from scientific papers, derived from arXiv preprints. This dataset, which includes raw LaTeXequations and their references, aims to facilitate the development of conversational models for interpreting scientific equations. We intend to make the dataset publicly available and establish baseline models for this purpose.

2 Related Work

Prior work in grounded question-answering has focused primarily on the visual modality (Antol et al., 2015; Tapaswi et al., 2016; Zhu et al., 2016; Lei et al., 2019; Singh et al., 2021). More recent work has focused on answering questions on grounded knowledge such as tables (Hannan et al., 2020; Nakamura et al., 2022; Sundar and Heck, 2023).

Simultaneously, the challenge of modeling mathematical equations has become an area of active research in Natural Language Processing. Chiang and Chen (2019) present an approach for modeling equations from math word problems. Wang et al. (2021a) present an approach for the dual problem of generating math word problems consistent with equations. Work on learning representations from equations includes MathBERT Peng et al. (2021), FORTE (Wang et al., 2021b), and Topiceq (Yasunaga and Lafferty, 2019). More recent work on generative architectures includes MathGPT (Scarlatos and Lan, 2023), an auto-regressive model based on GPT-2 (Radford et al., 2019) for various language+equation tasks. In contrast, CONVER-



Denote a translation invariant nearest-neighbour Hamiltonian as $\hat{H} = \sum_{n\in\mathbb{Z}} \hat{T}^n \hat{h}^n \hat{T}^{-n}$, where $\hat{T}$ is the shift operator and $\hat{h}$ acts non-trivially only on sites zero and one. We now try to approximate the time evolution generated by $\hat{H}$ of a uMPS $|\psi(A)\rangle$ without ever leaving the variational manifold of uMPS with fixed bond dimension D, by introducing a time-dependent parameterisation $A(t)$. Insertion into the time-dependent Schrödinger equation results in $\dot{A}^i \, |\partial_i \psi(A(t))\rangle = -\mathrm{i}\hat{H}\, |\psi(A(t))\rangle$, where we denote ∂_i for $\partial/\partial A^i$. Whereas the left hand side (LHS) is a linear combination of the tangent vectors $|\partial_i \psi(A(t))\rangle$ that span the tangent plane $T_A\mathcal{M}_{\mathrm{uMPS}}$, the right hand side (RHS) is a general vector in Hilbert space and this equation does not have an exact solution for $\dot{A}^i$. The best approximation is obtained by minimizing

where the argument $A(t)$ in every vector has been omitted for the sake of brevity. The LHS of Eq. (2) contains the $dD^2 \times dD^2$ Gram matrix of the tangent vectors $G_{\bar{i},j}(\overline{A}, A) = \langle \partial_{\bar{i}} \psi(\overline{A}) | \partial_j \psi(A) \rangle$. Expressions for this Gram matrix and the vector in the RHS of Eq. (2) are best derived using the explicit form for the tangent vector $B^i \, |\partial_i \psi(A)\rangle = \sum_{n\in\mathbb{Z}} \hat{T}^n \sum_{\{s_k\}=1}^{d} v_{\mathrm{L}}^{\dagger} \left(\cdots A^{s_{-1}} B^{s_0} A^{s_1} \cdots\right) v_{\mathrm{R}} \, |\mathbf{s}\rangle$, and are given by

The minimum is found by orthogonally projecting the evolution vector $\hat{H}\,|\psi(A(t))\rangle$ onto the tangent plane, as illustrated in Fig. 1. The resulting solution is determined by

$$\langle \partial_{\bar{j}} \psi | \partial_i \psi \rangle \, \dot{A}^i = -\mathrm{i} \, \langle \partial_{\bar{j}} \psi | \hat{H} | \psi \rangle \, , \qquad (2)$$

This construction can also be derived from an action principle and is known as the *time-dependent variational principle* (TDVP) [7, 8]. The resulting TDVP equations [Eq. (2)] can be shown to be sympletic [9]. Hence they respect energy conservation as well as conservation of all constants of motion, such as the expectation value of gen-

Constructing the relevant quantities and solving Eq. (2) for $\dot{A}^i$ involve operations with a computational complexity of $\mathcal{O}(D^6)$. Using an iterative method to implement $(1 - E)^{-1}$ and then solving for $\dot{A}^i$ can reduce this to $\mathcal{O}(D^3)$. However, the matrix $G_{\bar{i},j}$ is not invertible: because of the

```
{
  "dialogue": {
     “q1": “What is the motivation behind
introducing a time-dependent parameterization
for the uMPS?”,
     “a1": “The time-dependent parameterization
is introduced to approximate the time evolution
generated by the translation-invariant nearest-
neighbor Hamiltonian without leaving the
variational manifold of uMPS.”,
     “q2": “How is the solution for the time
evolution vector obtained in the context of the
time-dependent variational principle (TDVP)?”,
     “a2": “The solution involves minimizing the
norm of the difference between the left-hand
side and right-hand side of the time-dependent
Schrödinger equation. The minimum is found by
orthogonally projecting the evolution vector
onto the tangent plane of the uMPS manifold”,
     “q3": “What role does gauge invariance play
in the calculation involving the Gram matrix and
the null space of tangent vectors?”,
     “a3": “Gauge invariance leads to non-
invertibility of the Gram matrix due to linear
dependence in tangent vectors. Fixing the gauge
eliminates components in the null space,
resulting in a reduced-dimensional tangent
plane.”
   }
}
```

Figure 1: An example of the dataset construction process used in CONVERSATIONAL EQUATIONS

SATIONAL EQUATIONS is a dataset that addresses both grounded conversational question-answering and the modeling of mathematical equations.

3 Dataset Creation

3.1 Equation Extraction

We obtain grounded mathematical equations by parsing publicly available research papers published on arXiv [1], an open access repository of preprints of academic papers. Using AXCELL (Kardas et al., 2020), we obtain 15,000 LaTeXsource files from approximately 6,000 academic papers.

We parse the LaTeXfiles for all equation instances by searching for text inside the equation environment, that is, text within `\begin{equation}` and `\end{equation}` tags. To obtain relevant context regarding an equation in a self-supervised approach, we store the paragraph of text immediately before and after the equation. To obtain further context, we store all lines of text referring to the specific equation. In LaTeX, equations are often marked with a label (`\label{}`) for easy reference using the `\ref{}` command. Therefore, for each equation, we search for the label and if it exists, store all references that utilize this specific label. Using this approach, we obtain 42,500 equations in total across all documents.

3.2 Dialogue Generation

We prompt GPT-3.5 (Brown et al., 2020) to generate a sequence of three-turn question and answer pairs grounded on the equation and references and describe the process in Figure 1. Our prompt is:

I will give you an equation in latex form and a list of paragraphs which reference the equation. Given this information, I want you to generate three questions regarding the content, as well as the answers. Be brief and concise. Return the questions in JSON format like so: “dialogue”: {“q1": “question 1”, “a1": “answer 1”, “q2": “question 2”, “a2": “answer 2”, “q3": “question 3”, “a3": “answer 3”}.

4 Next Steps

This paper outlines ongoing work to collect the CONVERSATIONAL EQUATIONS dataset. Next steps include completing the collection of the dataset using the prompt-based approach. While the existing approach involves generating the entire conversation at once, we will also experiment with chain-of-thought prompting to generate inter-dependent conversational turns. For example, chain-of-thought could be used to detail parts of an equation sequentially building up to a final conversational turn that requires utilizing dialogue context to be answered. Along with the dataset, we will also release a baseline language model.

---

[1] `https://arxiv.org/`

Acknowledgements

This work was supported by NSF IIS-2112633 and by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ting-Rui Chiang and Yun-Nung Chen. 2019. Semantically-aligned equation generation for solving and reasoning math word problems. 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)*, pages 2656–2668, Minneapolis, Minnesota. Association for Computational Linguistics.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. *arXiv:2001.08034 [cs]*. ArXiv: 2001.08034.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. TVQA: Localized, Compositional Video Question Answering. *arXiv:1809.01696 [cs]*. ArXiv: 1809.01696.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv preprint arXiv:2105.00377*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alexander Scarlatos and Andrew Lan. 2023. Tree-based representation and generation of natural and mathematical language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3714–3730, Toronto, Canada. Association for Computational Linguistics.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal Input Multimodal Output Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online. Association for Computational Linguistics.

Anirudh Sundar and Larry Heck. 2022. Multimodal conversational AI: A survey of datasets and approaches. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Anirudh S. Sundar and Larry Heck. 2023. cTBLS: Augmenting large language models with conversational tables. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 59–70, Toronto, Canada. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. *arXiv:1512.02902 [cs]*. ArXiv: 1512.02902.

Zichao Wang, Andrew Lan, and Richard Baraniuk. 2021a. Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5986–5999, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zichao Wang, Mengxue Zhang, Richard G Baraniuk, and Andrew S Lan. 2021b. Scientific formula retrieval via tree embeddings. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1493–1503. IEEE.

Michihiro Yasunaga and John D Lafferty. 2019. Topiceq: A joint topic and mathematical equation model for scientific texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7394–7401.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. *arXiv:1511.03416 [cs]*. ArXiv: 1511.03416.