Augmenting Datasets for Fine-Tuning Large Language Models Using Semantic Variations

Alexander Chen[1,*], Caroline Tang[1] and Jennifer Sleeman[1]

[1]Johns Hopkins Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD 20723

Abstract

Fine-tuning Large Language Models (LLMs) to capture domain-specific knowledge remains constrained by the limited availability of high-quality, semantically rich training data, particularly in scientific domains where knowledge is distributed across heterogeneous textual sources. In this study, we explore a semantic variation methodology to augment training data by generating question-answer pairs with explicit control over semantic similarity. Rather than simple paraphrasing, this approach produces variations that expand the coverage of the underlying domain knowledge. We frame question-answer pairs as lightweight representations of relational knowledge, enabling their interpretation as implicit structures relevant to knowledge graph construction and semantic querying tasks. Using an Earth system science use case, we construct a dataset built from scientific literature and augment it with semantic variations generated by LLMs. We fine-tune the Meta LLaMA 3-8B model on these augmented datasets and compare performance against the base model. Additionally, we performed a sensitivity study in which we examined the effects of adjusting the number of variations injected into the base dataset and varying the question-answer pairs based on a semantic similarity score. This similarity score provides a way to control how closely each variation preserves the meaning of the original pair. We compared semantic variation datasets with different levels of similarity by fine-tuning LLaMA 3 on each dataset and measuring performance on a held-out set of questions. Our results show that semantically controlled augmentation improves domain-specific knowledge acquisition while preserving consistency, and we identify effective regimes for both the number of variations and the level of semantic similarity. These findings highlight the role of semantic-aware data augmentation in improving LLM performance in settings relevant to knowledge representation and semantic search.

Keywords

LLM, fine-tuning, question answering, semantic variation, data augmentation

1. Introduction

A central challenge in developing deep learning models is obtaining a dataset that is both sufficiently large and of sufficient quality for the target task. As a result, using models that have been pretrained on large, general-purpose datasets has enabled substantial progress in many areas of machine learning [1, 2, 3]. Large language models (LLMs), neural networks pretrained on massive text corpora to model and generate natural language [4], have specifically become increasingly prevalent due to their ability to capture rich semantic structure and broad knowledge that can be transferred across a wide variety of downstream tasks, including question answering, summarization, and classification [5]. This ability to capture and reuse semantic structure makes LLMs particularly well suited for tasks that involve representing and accessing domain knowledge. Pretrained models can either be used directly for downstream tasks, or they can be adapted to a specific domain through fine-tuning on a smaller task-specific dataset.

Fine-tuning LLMs remains useful in many settings where domain expertise, privacy, or cost constraints make direct reliance on large public models infeasible [6]. However, fine-tuning LLMs often requires datasets on the order of thousands of samples, to sufficiently adapt the model to a task of interest. For

ESWC'26: 5th Workshop on LLM-Integrated Knowledge Graph Generation from Text (TEXT2KG), May 10-14, 2026 | Dubrovnik, Croatia

*Corresponding author.
Alex.Chen@jhuapl.edu (A. Chen); Caroline.Tang@jhuapl.edu (C. Tang); Jennifer.Sleeman@jhuapl.edu (J. Sleeman)
0009-0008-5678-4218 (A. Chen); 0009-0003-9224-3666 (C. Tang); 0000-0002-8934-5587 (J. Sleeman)

many problems, obtaining datasets of this size is costly, if not impossible. As LLMs are increasingly used as AI assistants, their adoption in the physical and biological sciences has also grown [7, 8, 9], including applications in Earth system science. Despite this growing interest, the creation of fine-tuning datasets for scientific question answering remains a significant bottleneck [10]. Scientific knowledge tends to be distributed across a large and heterogeneous body of literature, and the construction of high-quality question-answer datasets that consistently represent this knowledge often requires substantial effort from domain experts.

In this work, we explored a methodology to address this challenge of building a large question-answer dataset to support an Earth system AI assistant. In our previous work [11, 12], we used a neuro-symbolic model to support an AI assistant for a question-answering task. In this previous work, natural language questions were translated into vectors that were used to invoke an AI discovery method, providing a mapping between unstructured queries and structured representations. The AI discovery method returned an output that was translated into a natural language response. In this early implementation of an AI assistant, the system was limited to answering questions that required yes/no responses.

In our current study, we expand the question-answering capabilities of this approach using a fine-tuned LLM to answer scientific questions beyond simple yes/no responses. We describe a method to overcome the burden of composing a dataset by using a human-AI approach to compose a scientifically-grounded question-answer dataset. We describe a new methodology for increasing the size of a question-answer dataset by creating *semantic variations* from a set of base question-answer pairs using an LLM and a semantic similarity metric, with the goal of preserving the underlying meaning of each pair. This approach enabled us to reach a dataset size sufficient for fine-tuning while maintaining semantic consistency across variations.

To study the utility and limitations of this approach, we performed a sensitivity study exploring how model performance changes as the similarity of semantic variations increases. We also compared the semantic variation augmentation approach against two baselines: one in which the dataset is augmented with exact duplicates of the original question-answer pairs, and another in which the dataset is augmented with completely random question-answer pairs. This sensitivity study illustrates how semantic variations can supplement a set of base question-answer pairs to produce dataset sizes suitable for fine-tuning LLMs while controlling for variation in meaning.

We first present results for fine-tuning on the original dataset curated by Earth system scientists. We then examine the impact on performance when augmenting the dataset with semantic variations. Finally, we present the results of the sensitivity study.

2. Background

The Earth system is composed of subsystems including the atmosphere, oceans, the cryosphere, and the land surface [13]. Scientists who study the Earth system seek to better understand the physical processes and feedback that governs interactions among these components. They examine how these systems behave over both short and long time horizons. The Earth system can be viewed as a nonlinear dynamical system that exhibits chaotic behavior [14]. As a result, multiple models are often run in parallel to improve predictive skill.

More recently, artificial intelligence (AI) has been applied to Earth system research for different tasks such as forecasting weather, climate, and air quality, as well as emulating physics-based numerical models and analyzing observational phenomena. As it relates to large foundation models, including LLMs, there have been a number of efforts to explore how these models can be applied to Earth system problems. This includes using LLMs for climate education [15], using observation-based foundation models to inform forecasting tasks [16, 17], and using foundation models trained on large Earth system simulation datasets as base models that are then fine-tuned for specific applications such as weather forecasting [18], air quality prediction [19], and related tasks. These efforts increasingly rely on integrating heterogeneous data sources and representations of Earth system knowledge.

In this work, we study the use of LLMs for answering scientific questions about a challenging Earth

system problem. In particular, our question-answering LLM translates scientific questions related to the potential collapse of the Atlantic Meridional Overturning Circulation (AMOC) into tool calls within our system, providing a structured interface between natural language queries and downstream analysis components. The AMOC is a major component of the global ocean circulation and has been identified as a potential tipping element in the climate system [20].

To support this translation, we fine-tune a common open-source LLM, the Meta LLaMA 3-8B model [21], which was chosen for its open-weight release and smaller parameter variants. This enables more practical fine-tuning compared to larger closed models while still providing strong performance across a broad range of language tasks.

Additionally, we used Low-Rank Adaptation (LoRA) [22], a fine-tuning technique that freezes the original model weights and trains only on lightweight, lower-rank matrices. LoRA was utilized in order to make the fine-tuning process easier and lighter, allowing us to train on fewer GPUs while also requiring less memory and enabling faster fine-tuning.

3. Related Work

Despite demonstrating impressive general abilities, LLMs may struggle with domain-specific scientific questions when precise terminology and evidence from the primary literature are necessary [23].

Retrieval-augmented generation (RAG) [24] can be used to address this knowledge gap, which grounds responses in external corpora, fetching supporting documents at inference time for models to cite and stay up-to-date. However, it remains vulnerable to potential retrieval errors, including irrelevant or missing evidence, and can suffer from degraded performance in long or noisy contexts [24]. These limitations motivate approaches that integrate some domain semantics into the model itself.

Fine-tuning complements RAG by internalizing stable domain knowledge, reducing dependence on potentially brittle retrieval. Parameter-efficient methods, such as Low-Rank Adaptation (LoRA) [22] and Quantization-Aware Low-Rank Adaptation (QA-LoRA) [25], make fine-tuning possible with little additional inference cost.

Fine-tuning can adapt models to specific domains [26], but can be challenging to perform as it is often dependent on the careful curation of high-quality datasets from subject matter experts (SMEs). Classic augmentation methods (e.g., Easy Data Augmentation, back-translation) increase lexical variety, mostly perturbing superficial form, but offer little control over meaning [27].

Related approaches in the semantic web and knowledge graph community have explored question answering through structured representations, including mapping natural language questions to formal queries such as SPARQL over domain knowledge graphs [28]. These methods provide strong guarantees on semantic precision and interpretability, but also often require task-specific training data that aligns natural language with structured query representations. Our work is complementary in that we focus on improving the quality and coverage of such training data through semantically controlled augmentation.

In contrast, LLM-based synthesis can create semantically guided examples and has been increasingly studied as a means of generating synthetic instruction and training data at scale [29, 30, 31]. For example, Self-Instruct bootstrap instruction data [29] from model generations and "Textbooks Are All You Need" [30] show that curated synthetic material can be effective. In scientific Q&A specifically, SME-constructed corpora such as PubMedQA [32], BioASQ [33], and SciFact [34] are high quality, but costly to scale, which makes our strategy of starting from a base dataset of SME question-answer pairs and then using an LLM to generate semantic variations a natural way to expand coverage without deviating from domain truth.

Recent work has also explored domain-specific adaptation of LLMs for structured and specialized tasks, such as information extraction and classification in scientific and financial domains [35], highlighting the importance of preserving semantic fidelity during data augmentation.

By combining human curation with LLM assistance, this synthesis approach can significantly reduce labor-intensive work that would otherwise require multiple scientists to carefully curate examples. We show how to achieve this using the Earth system problem of AMOC collapse as a use case.

4. Enriching Fine-Tuning Datasets with Semantic Variations

In this section, we present a semantic variation framework for augmenting an existing dataset for Earth system question answering. We analyze the resulting semantically enriched dataset and evaluate its utility by using it to fine-tune LLaMA 3 and assessing the performance of the resulting fine-tuned model. We use GPT-4 [36] for consistency with our prior pipeline and to maintain comparability across experiments. However, more recent models, such as GPT-5, may further improve the quality of semantic variations, which we leave for future work.

4.1. The Semantic Variation Methodology

We use a *semantic variation* methodology to expand an existing dataset of question-answer pairs by prompting an LLM, in this case GPT-4, to generate semantically similar yet distinct versions of each base question-answer pair. We focus on questions with short responses that are suitable for fine-tuning LLMs, although the overall approach may be applicable to a broader range of domain adaptation settings. Let the original dataset be

$$\mathcal{D}_0 = \{(q_i, a_i)\}_{i=1}^{N},$$

where q_i denotes a base question, a_i its corresponding answer, and N the number of original examples. For each base pair (q_i, a_i), we use an LLM-based generator G_ϕ to produce K semantic variations,

$$G_\phi(q_i, a_i; r) \to (\tilde{q}_{i,r}, \tilde{a}_{i,r}), \qquad r = 1, \ldots, K.$$

The resulting augmented dataset can then be defined as

$$\mathcal{D}_{\text{sem}}^{(K)} = \mathcal{D}_0 \cup \left\{ (\tilde{q}_{i,r}, \tilde{a}_{i,r}) : i = 1, \ldots, N,\ r = 1, \ldots, K \right\},$$

resulting in a total size of $N(1 + K)$. This procedure enables datasets containing hundreds of expert-curated examples to be expanded to thousands of training instances.

A key requirement of our *semantic variation* method is that each generated pair preserves the underlying scientific meaning and relationships of the original example while varying only its linguistic form. In our approach, we enforce this through a semantic similarity constraint. For each generated variation, $(\tilde{q}_{i,r}, \tilde{a}_{i,r})$, we compute a semantic similarity score relative to the original pair (q_i, a_i) using a chosen similarity metric, $S(\cdot, \cdot)$. A variation is retained only if

$$S\big((q_i, a_i), (\tilde{q}_{i,r}, \tilde{a}_{i,r})\big) \geq \tau,$$

where τ is a predefined similarity threshold.

4.2. Building the Base Dataset

To build the base dataset we first collected a domain-specific corpus based on scientific papers found by querying Web of Science on February 15, 2024 using the words "AMOC" and "slow*". The search returned 116 papers, spanning 2000-2024, including 85 papers from 2000-2020 and 31 papers from 2021-2024, drawn from major ocean and climate science publishers and society journals, including AGU/Wiley, EGU/Copernicus, AAAS/Science, Springer Nature/Springer, Elsevier, AMS, and related venues. For each selected article, we retrieved and standardized the full text, segmented the content into sections (abstract, introduction, methods, results, and discussion/conclusion) and performed a manual check to ensure extraction fidelity and section alignment.

We then generated section-grounded question-answer pairs using a GPT-4-based pipeline. For each paper section, the model was prompted to produce five question-answer pairs, yielding 25 pairs per article. To encourage explanatory and quantitative reasoning, we constrained the distribution so that no more than half were binary (yes/no, or true/false). SMEs then reviewed and edited each question-answer pair using a three-level rubric: (1) coherent and answerable from the associated section; (2) section-relevant but requiring revision for accuracy/wording; and (3) incoherent or not grounded in the

section. Only SME-validated question-answer pairs were retained in the base dataset. This manually reviewed base dataset then served as the foundation for the controlled semantic variation procedure used to expand the fine-tuning corpus while preserving scientific grounding in AMOC-related literature.

4.3. Measuring Similarity Between Semantic Variations and Original Questions

Using the base dataset of question-answer pairs, GPT-4 was then used to generate semantic variations of each pair.

We flagged questions as True/False or short answer to ensure that the newly generated answers would be in the desired format. To assess the effects of increasing the number of semantic variations in a dataset, we generated augmented datasets containing 5, 10, 15, and 20 total versions of each question-answer pair, including the original.

For fewer variations, all variants could be generated jointly in a single prompt. For larger variations, token constraints required sequential generation. At each sequential iteration, previous variations were appended to the prompt, giving the LLM additional context of existing variations that it should avoid.

We evaluated the generated semantic variations to determine whether they introduced linguistic diversity while preserving the meaning of the original questions. To do this, we used both ROUGE-1 and SBERT (Sentence-BERT) [37]. ROUGE-1 is a unigram-overlap metric that measures recall. Given an original question x and a generated variation y, ROUGE-1 recall is defined as

$$\text{ROUGE-1}(x, y) = \frac{\sum_{w \in x} \min\big(\text{count}_x(w), \text{count}_y(w)\big)}{\sum_{w \in x} \text{count}_x(w)}$$

where $\text{count}_x(w)$ and $\text{count}_y(w)$ denote the number of times unigram w appears in x and y, respectively, capturing lexical overlap. To assess semantic similarity, we used an SBERT cross-encoder, which jointly processes the sentence pair and outputs a similarity score,

$$\text{SBERT}(x, y) = f_\theta(x, y),$$

where f_θ denotes the cross-encoder model.

Using these two metrics together allowed us to assess similarity from both a lexical and semantic perspective as a proxy for preserving the underlying meaning of the represented knowledge. The resulting distributions are shown in Figures 1a and 1b. The similarity was assessed by comparing the variation with the original question from which it was generated.

The results indicate that increasing the size of the dataset by increasing the number of semantic variations per original question had little effect in terms of the ROUGE-1 score. No variation showed a significantly higher ROUGE-1 score. However, increasing the number of semantic variations did influence the SBERT score. As the number of variations increased, the score tended to improve.

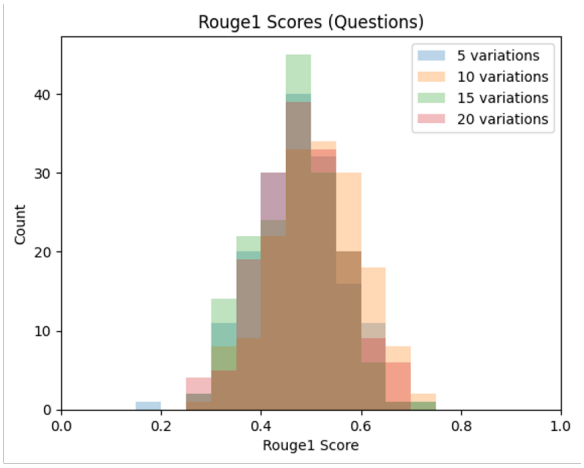
The relatively stable ROUGE-1 scores suggest that increasing the number of generated variations does not produce progressively more lexically repetitive variants of the original questions. Additionally, consistently high SBERT scores indicate that most generated variations remained semantically close to their source questions.

We note that while semantic similarity metrics provide a practical way to control and evaluate variation quality, they do not directly measure factual correctness, which remains an important consideration for scientific question answering.

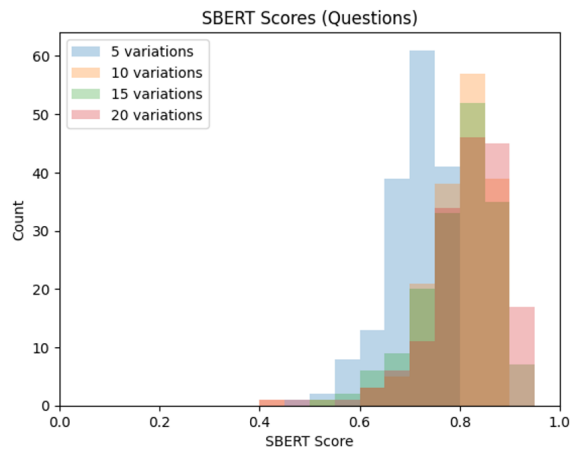
Having established that the generated variations preserved semantic content while introducing linguistic diversity, we then evaluated the effects on downstream fine-tuning using the augmented dataset.

4.4. Fine-Tuning Experiments

Using the 20-version dataset, the total set of question-answer pairs increased from a size of 399 pairs in the initial dataset to a total size of 7980 pairs. A train, validation, and test split resulted in a training



(a) ROUGE-1 scores for measuring lexical overlap comparing semantic variation questions with the original base question.



(b) SBERT scores for measuring semantic similarity by comparing semantic variation questions with the original base question.

Figure 1: Scores from measuring both lexical overlap using ROUGE-1 (a) and semantic similarity using SBERT (b).

dataset size of 4680 pairs, a validation set size of 1660 pairs, and a test set size of 1640 pairs. To prevent data leakage, the variations were grouped with their original question. The full set of semantic variations was generated in less than one day.

For fine-tuning, each question-answer pair was appended with the source from which it had been derived as contextual input. We used a structured prompt template with explicit markers that identify the context, question, and answer fields. The template also specified the expected response format, such as a single-word answer for binary questions or a short paragraph for short-answer questions. During the evaluation, the same prompting structure was used, but with the answer field omitted. The base prompt is shown in Figure 2. This prompt was given to the model for each question-answer pair during fine-tuning, with the appropriate context, question, and answer attached directly afterwards.

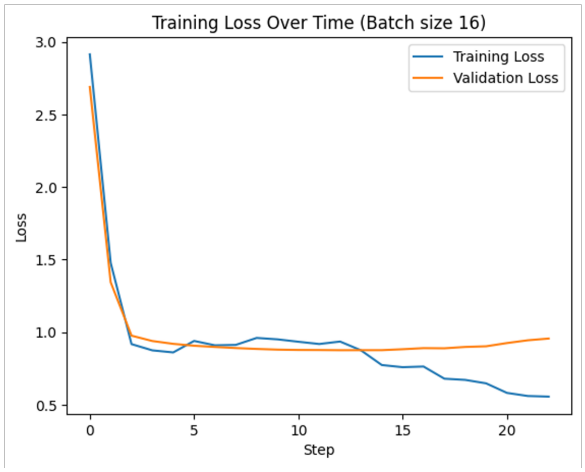
```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a helpful, respectful and honest assistant. Always answer as
   helpfully as possible, while being safe. Answer a question one time and one time only.
   If a question does not make any sense, or is not factually coherent,
   explain why instead of answering something not correct. If you don't know the answer to
   a question, please don't share false information. For each question, utilize the context
   provided to give your answer. DO NOT REPEAT THE QUESTION. If you see [QUESTION: TRUE/FALSE],
   answer with either only TRUE or FALSE. TRUE means you agree with the stated question whereas
   FALSE means that it is incorrect. Do not answer with anything else and provide an explanation.
   If you see [QUESTION: SHORT ANSWER], answer with a short paragraph, explaining your reasoning
   from the text.
   \n<|eot_id|><|start_header_id|>user<|end_header_id|>\n"""
```

Figure 2: An example of a prompt used during the fine-tuning process. It provides specific instructions as to how to recognize and handle each type of question.

Throughout the fine-tuning process, we tuned several hyperparameters. Specifically, we varied the LoRA scaling factor α, the LoRA rank r, and the batch size. The LoRA scaling factor α determines the magnitude of the low-rank updates applied to the frozen base weights, and the LoRA rank r controls the dimensionality of the low-rank update, with higher values allowing the model to capture more task-specific adjustments along with a higher number of trainable parameters. We started with an initial setup of $\alpha = 16$, $r = 64$, and a batch size of 4. As shown in Figure 3a, performance was poor using the base hyperparameter values. The validation loss increased, while the training loss decreased.

We then performed a hyperparameter sweep over different (α, r) pairs and observed that performance improved when smaller values of α were paired with larger values of r. Because these two parameters jointly determine the behavior of the LoRA update, we evaluated how different combinations performed. This trend suggests that better results were obtained when the LoRA adapters had greater

(a) Fine-tuning LLaMA 3 using LoRA and the base hyperparameter setup. Loss plot shows model is unable to generalize to the validation dataset.



(b) Fine-tuning LLaMA 3 using LoRA and a hyperparameter setup based on hyperparameter space exploration. Loss plot shows model is able to generalize to the validation dataset

Figure 3: Comparing a fine-tuned LLaMA 3 using LoRA with the base hyperparameters (a) and improved hyperparameters (b).

representational capacity but applied their updates more conservatively. We also independently varied batch size and observed improvements in performance when batch sizes were larger. In the end, a final setup with $\alpha = 4$, $r = 128$, and batch size of 16 was used. As shown in Figure 3b, the validation loss improved significantly. We used these settings to evaluate the performance of this model. Fine-tuning required two H100 GPUs and approximately eight hours to complete.

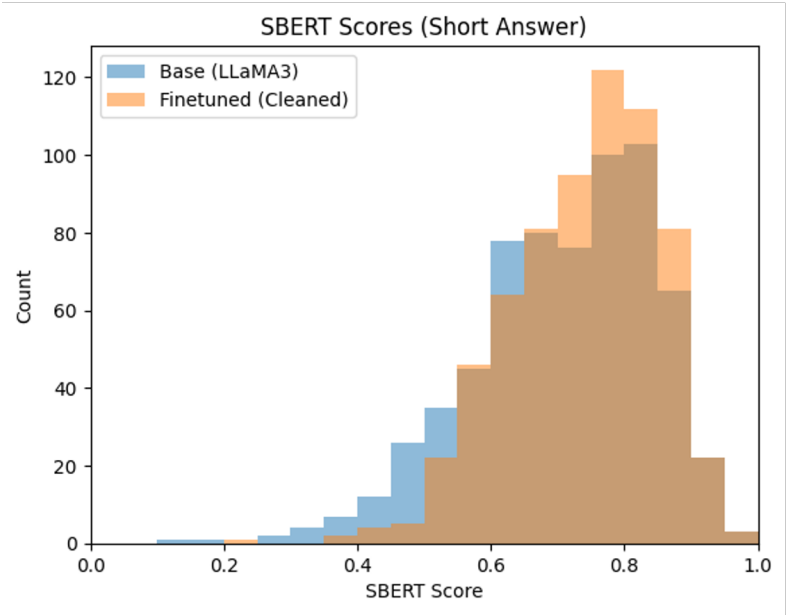


Figure 4: SBERT Scores - Comparing the fine-tuned LLaMA 3 using semantic variations with the base LLaMA 3.

4.5. Results

Model performance was evaluated primarily using SBERT to quantify the semantic similarity between each generated answer and the corresponding ground-truth answer in the test set. Before evaluating the output of LLaMA 3, the output went through a post-processing step. Using a heuristic, we removed

artifacts introduced by the prompt template, such as residual flags.

In Figure 4, we compare the SBERT score using LLaMA 3 out of the box and the fine-tuned LLaMA 3 using our semantically varied dataset. Given this comparison, it appears that the fine-tuned model shows some improvement in SBERT semantic similarity over the base model, suggesting that the fine-tuned model better captures the underlying domain knowledge represented in the training data.

5. Semantic Variations - A Sensitivity Study

To better understand the limitations of the *semantic variation* methodology, we conducted a sensitivity study focused on similarity as a metric for creating semantic variations. Specifically, we partitioned question-answer semantic variations into different bins of data based on their SBERT similarity to the original question. We then fine-tuned and evaluated separate models on datasets constructed from each similarity bin using a consistently held-out test set.

5.1. Experimental Setup

We limited our dataset to only short-answer questions for this sensitivity study, resulting in a base dataset of 165 question-answer pairs. The dataset was then split into 100 questions for the training set, 15 for the validation set, and 50 for the test set. The data were binned according to the similarity ranges defined in Table 1.

Table 1
SBERT was used to create bins based on similarity between a semantic variation and its original base question.

Bin Type	Similarity Range
Low Similarity	12.5%-37.5%
Medium Similarity	>37.5%-62.5%
High Similarity	>62.5%-87.5%
Duplication	100%
Random	Unknown

We chose these low, medium, and high ranges specifically to exclude question-answer pairs that were too distant from the base question-answer pair as well as those that were near duplicates. Additionally, we built two datasets: one with exact duplicates of the original set of questions and one with a completely random set of questions, sourced from a database of Jeopardy! questions [38].

These datasets were created as a means to test the extreme boundaries of semantic variations, from both no variation at all to entirely semantically different examples. In total, there were five distinct datasets that represented separate ranges of similarity. For the sensitivity study, we also increased the number of generated variations to 29 variations of each base question-answer pair for a total of 30 versions. Each binned dataset resulted in a training set of 3000 questions and a validation set of 450 questions. The same base test set was used for all binned datasets. Examples of these binned semantic variations are shown in Figure 5.

Fine-tuning was performed for all five ranges using the same hyperparameter setup as in the previous model. The loss plots for the five models are shown in Figure 6. Training loss plots indicate relative stability for each model except the one that included exact duplicates of the questions. In this model, the validation loss increases after 2.5 steps and does not recover.

5.2. Results

With the fine-tuned models, we examined the differences in their resulting performance. We compared the answer of each model for a specific question with the correct associated answer from the data set using SBERT. All five models were tested on the same test set of 50 question-answer pairs from the original base dataset. The test set was kept completely separate from the training and validation sets.

Original Question	Random	Low Similarity	Med Similarity	High Similarity	Exact Duplicate
What methodology was used to estimate heat transport and freshwater convergence?	Who appointed George C. Marshall as Secretary of Defense?	How is climate change affecting ocean currents?	How was the movement of warm water and salt content measured in the marine environment?	How did researchers calculate the movement of heat and the gathering of freshwater?	What methodology was used to estimate heat transport and freshwater convergence?
What was investigated in terms of the largest contributor to anomalies in the South Atlantic heat budget?	What is the name of the knife type that includes a Ranger model featuring a wood saw, can opener, and, of course, a toothpick?	In the context of oceanography, which current is significant for the transfer of warm water into the South Atlantic?	What aspect of oceanic circulation affects thermal irregularities in the South Atlantic Ocean?	What was the focus of research regarding the primary factor affecting irregularities in the South Atlantic's thermal balance?	What was investigated in terms of the largest contributor to anomalies in the South Atlantic heat budget?
What is considered a necessary step before assessing the performance of decadal predictions?	What adjective, when combined with ""glory,"" forms a word meaning unwarranted pride?	Why is understanding historical climate patterns important for future projections?	What should be examined to ensure the accuracy of long-term weather forecasts?	What must be done prior to evaluating the success of ten-year forecasts?	What is considered a necessary step before assessing the performance of decadal predictions?
Why is understanding the Labrador Sea important for predicting decadal changes in the North Atlantic?	Who was the British baron that traveled to Bretton Woods, New Hampshire, in 1944, where the IMF was created?	How does the Atlantic Meridional Overturning Circulation affect climate patterns across Europe?	What role does the formation of deep water in the Labrador Sea play in oceanic currents?	How does the Labrador Sea impact the prediction of decade-scale climate variations in the North Atlantic region?	Why is understanding the Labrador Sea important for predicting decadal changes in the North Atlantic?
How do flux adjustments affect the near surface salinity and Fov in the model?	What state and its state capital both have names that end with the letters "na"?	What is the impact of altering freshwater inputs on ocean circulation patterns?	What impact do adjustments in flux have on oceanic thermal layers and their stability?	What impact do adjustments in flux have on the model's shallow water salinity and its fresh water balance?	How do flux adjustments affect the near surface salinity and Fov in the model?

Figure 5: An example set of questions that shows the original base question and an example semantic variation from each bin.

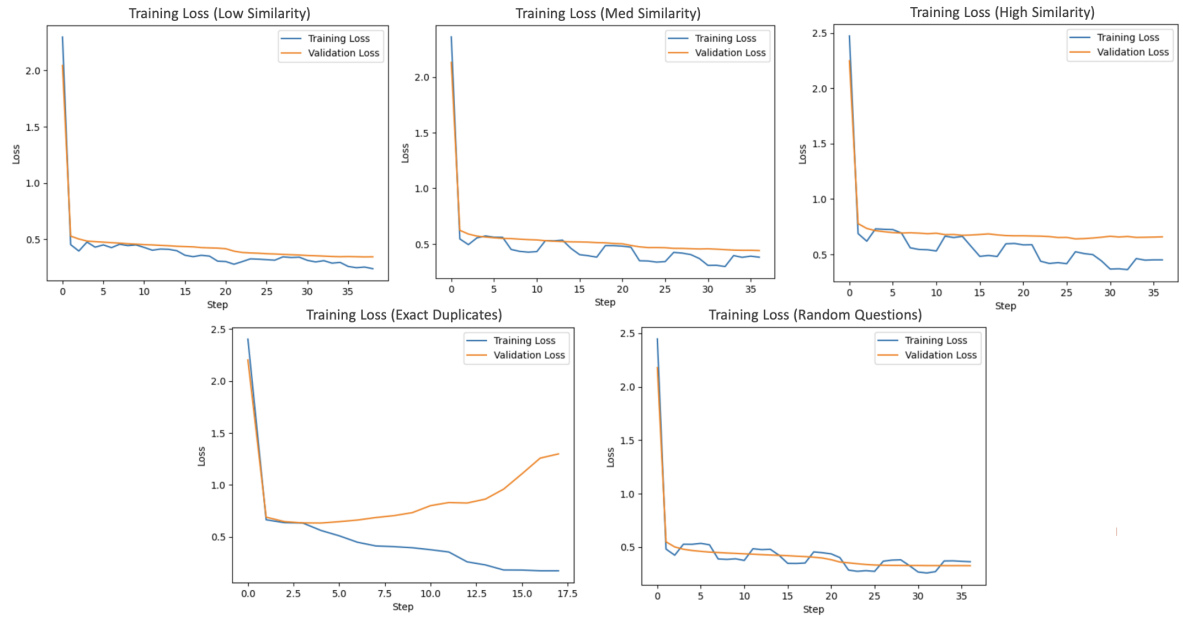


Figure 6: Training Loss Plots for each fine-tuned model across all binned semantic variation datasets after applying hyperparameter tuning.

The model output was cleaned using the previously described heuristic, and the SBERT results are shown in Figure 7.

Comparing the results of the models fine-tuned on different training datasets, we saw that the model trained on the Jeopardy! dataset performed worst. Of the models trained on semantically enriched datasets, those trained on medium- and high-similarity datasets performed the best, having the least variability in SBERT scores across the test set while maintaining high average scores. The model trained on the low-similarity dataset seemed to perform similarly to the model trained on exact duplicates, with both having a wider range of scores including outliers in which the model performed poorly on some examples from the test set.

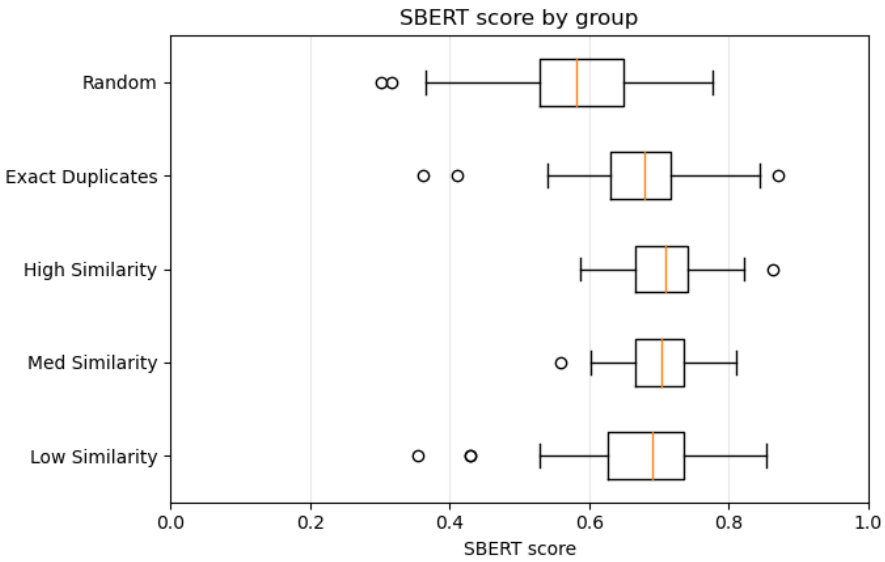


Figure 7: Sensitivity study using multiple semantic variation bins, tested on only original questions. Measuring performance using SBERT, medium- and high-similarity datasets appear to offer the best performance when comparing LLM answers with the held-out answer.

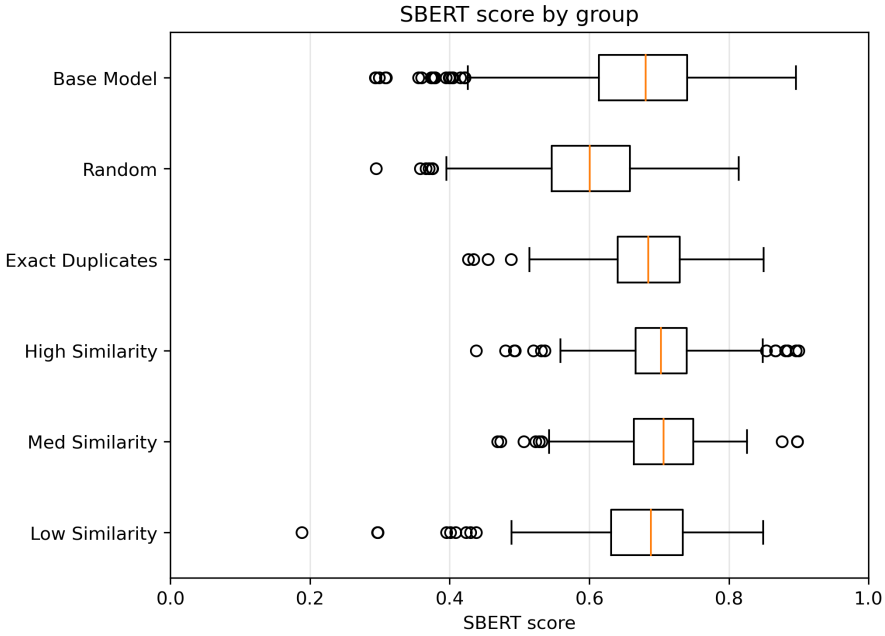


Figure 8: Sensitivity study using multiple semantic variation bins and comparing with the original base model, tested on original questions and variations. Measuring performance using SBERT, models fine-tuned on medium- and high-similarity datasets perform similarly and appear to offer the best performance, outperforming the base model when comparing LLM answers with the held-out answer along with tighter quartile tails, suggesting higher reliability.

We also tested these models against the original test set used in the experiment shown in Figure 4. We removed questions from our original test set that the current models had seen during training and validation, as well as any variations related to these base questions to prevent data leakage. These results are presented in Figure 8. Among the semantically enriched datasets, the models trained on medium- and high-similarity variations performed best. In both cases, the median of the score distribution exceeded that of the base model with a smaller spread, suggesting not only improved performance but also greater consistency. These fine-tuned models outperformed the base model, and the general trends

across all binned models were consistent with the sensitivity results. The model fine-tuned on random Jeopardy! question-answer pairs performed the worst. Surprisingly, the exact duplicates model did not perform much worse than the high-similarity model.

5.3. Limitations

One limitation in our evaluation of this data augmentation approach stems from our use of LoRA as the fine-tuning method. LoRA was initially selected due to computational resource constraints, as it provides a lightweight and efficient alternative to full model fine-tuning. However, because LoRA keeps most of the base model weights frozen and only learns low-rank adapter updates, the resulting parameter updates are relatively constrained. As a result, the performance gains from fine-tuning were likely smaller than what might be achieved with full model fine-tuning, where the underlying model weights are allowed to adapt more extensively to the augmented dataset.

Additionally, our evaluation relies on a semantic similarity metric (SBERT), which measures closeness in meaning but does not directly assess factual correctness. This is an important limitation for scientific question answering, where accuracy and faithfulness to the underlying literature are critical. While SBERT provides a useful proxy for evaluating semantic alignment between generated and reference answers, future work should incorporate evaluation methods that explicitly account for factual correctness and scientific validity.

6. Conclusions and Future Work

In this work, we describe an approach for constructing a large scientific dataset from the scientific literature, along with a new *semantic variation* methodology for expanding a dataset to sizes suitable for fine-tuning LLMs. We evaluate datasets generated with different levels of semantic variation in order to better understand how dataset size and augmentation affect the performance of the fine-tuned model. Although advances in LLM capabilities increasingly reduce the need for fine-tuning, there remain important cases where domain adaptation through fine-tuning is still beneficial.

Our results suggest that semantic variation offers a practical strategy for overcoming one of the primary barriers to fine-tuning: the difficulty of constructing sufficiently large, high-quality training datasets. More broadly, this approach highlights the importance of preserving semantic consistency when expanding datasets that represent domain knowledge, which is particularly relevant for applications involving knowledge representation, semantic search, and question answering over structured or semi-structured data.

Additionally, the generation of variations was done using the GPT-4 model, which was chosen primarily because it was one of the leaders at the time this research was conducted, and also because of its ease of use. In future work, we plan to explore the use of more recent LLMs to generate semantic variations, as well as evaluation strategies that better capture factual correctness in scientific question answering. Expanding this approach to other areas besides Earth science would also allow us to compare against other state-of-the-art models such as PubMedQA and SciFact, something not possible in this iteration of the work due to differing domains. We also aim to investigate how semantically controlled augmentation can support downstream tasks such as knowledge extraction and text-to-knowledge graph pipelines.

Acknowledgments

The authors thank the Earth system researchers who helped curate the base dataset used in this research, including Dr. Meghan L. Troup, Dr. Genevieve Jay Brett, Dr. Jennifer M. Boothby, Jessica K. Makowski, and Jeremy J. Bruch.
This work was supported by independent research and development funding from the Research and Exploratory Development Mission Area of the Johns Hopkins Applied Physics Laboratory.

Declaration on Generative AI

During the preparation of this work, the author(s) used **ChatGPT** in order to: refine the English language, correct grammar, and improve readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Communications of the ACM 60 (2017) 84–90.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[3] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in: Advances in Neural Information Processing Systems, volume 33, 2020.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[6] X.-K. Wu, M. Chen, W. Li, R. Wang, L. Lu, J. Liu, K. Hwang, Y. Hao, Y. Pan, Q. Meng, et al., LLM Fine-tuning: Concepts, Opportunities, and Challenges, Big Data and Cognitive Computing 9 (2025) 87.

[7] S. Schmidgall, Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, Z. Liu, E. Barsoum, Agent Laboratory: Using LLM Agents as Research Assistants, arXiv preprint arXiv:2501.04227 (2025).

[8] S. Ren, P. Jian, Z. Ren, C. Leng, C. Xie, J. Zhang, Towards Scientific Intelligence: A Survey of LLM-based Scientific Agents, arXiv preprint arXiv:2503.24047 (2025).

[9] D. Pantiukhin, B. Shapkin, I. Kuznetsov, A. A. Jost, N. Koldunov, Accelerating Earth Science Discovery via Multi-Agent LLM Systems, Frontiers in Artificial Intelligence 8 (2025) 1674927.

[10] T. Saikh, T. Ghosal, A. Mittal, A. Ekbal, P. Bhattacharyya, ScienceQA: a novel resource for question answering on scholarly articles, International Journal on Digital Libraries 23 (2022) 289–301. doi:10.1007/s00799-022-00329-y.

[11] J. Sleeman, D. Chung, C. Ashcraft, J. Brett, A. Gnanadesikan, Y. Kevrekidis, M. Hughes, T. Haine, M.-A. Pradal, R. Gelderloos, et al., Using Artificial Intelligence to aid Scientific Discovery of Climate Tipping Points, arXiv preprint arXiv:2302.06852 (2023).

[12] C. Ashcraft, J. Sleeman, C. Tang, J. Brett, A. Gnanadesikan, Neuro-Symbolic Bi-Directional Translation - Deep Learning Explainability for Climate Tipping Point Research, 2023. URL: https://arxiv.org/abs/2306.11161. doi:10.48550/arXiv.2306.11161. arXiv:2306.11161.

[13] G. M. Flato, Earth system models: an overview, Wiley Interdisciplinary Reviews: Climate Change 2 (2011) 783–800.

[14] H. A. Dijkstra, Nonlinear Climate Dynamics, Cambridge University Press, 2013.

[15] D. Thulke, Y. Gao, P. Pelser, R. Brune, R. Jalota, F. Fok, M. Ramos, I. van Wyk, A. Nasir, H. Goldstein, T. Tragemann, K. Nguyen, A. Fowler, A. Stanco, J. Gabriel, J. Taylor, D. Moro, E. Tsymbalov, J. de Waal, E. Matusov, M. Yaghi, M. Shihadah, H. Ney, C. Dugast, J. Dotan, D. Erasmus, ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change, 2024. URL: https://arxiv.org/abs/2401.09646. arXiv:2401.09646, arXiv:2401.09646.

[16] D. Szwarcman, S. Roy, P. Fraccaro, O. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. De Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, et al., Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications, IEEE Transactions on Geoscience and Remote Sensing (2025).

[17] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, et al., TerraMind: Large-Scale Generative Multimodality for Earth Observation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 7383–7394.

[18] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, et al., Learning skillful medium-range global weather forecasting, Science 382 (2023) 1416–1421.

[19] C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, J. Brandstetter, P. Garvan, M. Riechert, J. Weyn, H. Dong, A. Vaughan, et al., Aurora: A Foundation Model for the Earth System, arXiv preprint arXiv:2405.13063 1 (2024).

[20] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, H. J. Schellnhuber, Tipping elements in the Earth's climate system, Proceedings of the national Academy of Sciences 105 (2008) 1786–1793.

[21] A. Grattafiori, A. Dubey, A. Jauhri, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783 (2024). doi:10.48550/arXiv.2407.21783.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv:2106.09685 (2021).

[23] D. Duong, B. D. Solomon, Analysis of large-language model versus human performance for genetics questions, European Journal of Human Genetics 32 (2024) 466–468. doi:10.1038/s41431-023-01396-8.

[24] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, et al., Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv preprint arXiv:2312.10997 2 (2023) 32.

[25] Y. Xu, L. Xie, X. Gu, X. Chen, H. Chang, H. Zhang, Z. Chen, X. Zhang, Q. Tian, QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models, arXiv preprint arXiv:2309.14717 (2023).

[26] H. Tran, Z. Yang, Z. Yao, H. Yu, BioInstruct: Instruction Tuning of Large Language Models for Biomedical Natural Language Processing, Journal of the American Medical Informatics Association 31 (2024) 1821–1832. doi:10.1093/jamia/ocae122.

[27] J. Chen, D. Tam, C. Raffel, M. Bansal, D. Yang, An Empirical Survey of Data Augmentation for Limited Data Learning in NLP, Transactions of the Association for Computational Linguistics 11 (2023) 191–211. doi:10.1162/tacl_a_00542.

[28] J. C. Rangel, T. M. de Farias, A. C. Sima, N. Kobayashi, Sparql generation: an analysis on fine-tuning openllama for question answering over a life science knowledge graph, arXiv preprint arXiv:2402.04627 (2024).

[29] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-Instruct: Aligning Language Models with Self-Generated Instructions, in: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), 2023, pp. 13484–13508.

[30] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, Y. Li, Textbooks Are All You Need, arXiv:2306.11644 (2023).

[31] O. Honovich, T. Scialom, O. Levy, T. Schick, Unnatural instructions: Tuning language models with (almost) no human labor, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 14409–14428.

[32] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, X. Lu, PubMedQA: A Dataset for Biomedical Research Question Answering, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2567–2577. URL:

https://aclanthology.org/D19-1259/. doi:10.18653/v1/D19-1259.
[33] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artiéres, A.-C. Ngonga Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (2015) 138. URL: https://doi.org/10.1186/s12859-015-0564-6. doi:10.1186/s12859-015-0564-6.
[34] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or Fiction: Verifying Scientific Claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609/. doi:10.18653/v1/2020.emnlp-main.609.
[35] M. Birti, A. Maurino, F. Osborne, Optimizing Large Language Models for ESG Activity Detection in Financial Texts, in: Proceedings of the 6th ACM International Conference on AI in Finance, 2025, pp. 856–863.
[36] OpenAI, ChatGPT, https://chat.openai.com/chat, 2023.
[37] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv preprint arXiv:1908.10084 (2019). doi:10.48550/arXiv.1908.10084.
[38] Aravind Ram Nathan, Jeopardy Dataset, https://www.kaggle.com/datasets/aravindram11/jeopardy-dataset-updated, 2021. Kaggle dataset.

Online Resources

The sources for this work are available via GitHub at: https://github.com/JHUAPL-ACTS/Semantic-Variation-Finetuning.