ArgSciChat: A Dataset of Argumentative Dialogues on Scientific Papers

Federico Ruggeri1 * Mohsen Mesgar2 Iryna Gurevych2

1 DISI, University of Bologna
2 Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science
Technical University of Darmstadt
1 federico.ruggeri6@unibo.it
2 www.ukp.tu-darmstadt.de

Abstract

With recent advances in question-answering models, various datasets have been collected to improve and study the effectiveness of these models on scientific texts. Questions and answers in these datasets aim to explore the content of a scientific paper by seeking information from the paper’s content. However, these datasets do not tackle the argumentative content of scientific papers, which is of huge importance in persuasiveness of a scientific discussion. We introduce ArgSciChat, a dataset of 41 argumentative dialogues between scientists on 20 NLP papers. The unique property of our dataset is that it includes both exploratory and argumentative interactions in a dialogue on a scientific paper. Moreover, the size of ArgSciChat demonstrates the challenges in collecting dialogue for specialized domains. Thus, our dataset is an aid for evaluating few-shot transfer learning methods for dialogue agents in a specialized domain. We perform a detailed analysis on ArgSciChat and use it to fine-tune and evaluate a pre-trained document-grounded dialogue agent for NLP papers. The results show that dialogues in ArgSciChat include exploratory and argumentative interactions, which makes it challenging for the examined dialogue agent to answer questions. We publicly release ArgSciChat.1

1 Introduction

The enormous and ever-growing number of scientific papers (Munroe, 2013; Ronzano and Saggion, 2015) make scientific text processing imperative. A considerable body of research in NLP is dedicated to developing methods to provide insights from scientific papers (Beltagy et al., 2019; Zhang et al., 2021; Wadden et al., 2022; Zhang et al., 2021; Parveen et al., 2016; Wadden et al., 2022; Mysore et al., 2022). Recent advances in Question-Answering (QA) models made them a great asset in accessing the content of scientific papers (Dasigi et al., 2021). To develop and evaluate QA models on scientific papers, various datasets have been collected (Dasigi et al., 2021; Pampari et al., 2018; Pappas et al., 2018; Jin et al., 2019; Tsatsaronis et al., 2015; Krallinger et al., 2020). For each scientific paper in these datasets, a set of questions with the exploratory (EXP) intention is defined. These questions aim to explore the content of a scientific paper and seek information. Answers to EXP questions are limited to single words or short text spans extracted from the text of the paper.

Scientific answers should ideally be explained with argumentation to be persuasive. If an answer is not persuasive, then a follow-up question might be raised to argue about the answer. This property of scientific domain motivates a crucial need for a dataset of scientific argumentative dialogues on scientific papers. Collecting such a dataset is challenging for two reasons. First, while there are linguistic theories about scientific argumentative dialogues (Walton, 2008; Macagno, 2000); there is no functional intent sets for collecting argumentative questions and answers. Second, scientists (and experts in general) are on a strictly limited time-budget, so incentivizing scientists and aligning them to chat about a scientific paper should be very optimized. Although off-the-shelf crowdsourcing platforms are applicable for data collection in generic domains, they do not lend themselves for collecting argumentative dialogues in science.

In this work, we introduce ArgSciChat, the first dataset of annotated argumentative dialogues on NLP papers. Figure 1 shows an example dialogue from our dataset. To collect ArgSciChat, we first define a set of intent classes for interactions between scientists about scientific papers. We then develop a new dialogue collection methodology in
which we involve scientists preferences into the dialogue collection process. We allow scientists to choose a paper to chat about based on their preference. Then align them with each other to conduct a dialogue about their chosen papers according to their schedule. In ArgSciChat, the dialogues are in English and conducted synchronously by subjects who are NLP scientists. Each dialogue is about a single paper. We define the aim of dialogues to exchange information and argue about retrieved answers from the content of a paper. Each subject plays a different role in the dialogue. E is the expert on the research presented in the paper, and P (for proponent) is the subject aiming to learn about the scientific paper by posing questions to the expert. Importantly, E and P can argue about the content of the paper and each other’s opinion.

We extensively analyze the properties of the collected dialogues in ArgSciChat. To put our dataset in context, we compare our dataset with CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018) and Doc2Dial (Feng et al., 2020). We then evaluate the diversity of dialogues in our dataset and analyze the quality of answers. Finally, we use our dataset to fine-tune and evaluate LED (Beltagy et al., 2020), a generic dialogue agent for answering questions about a long text. The goal of this experiment is to study the difficulty level of our dataset for current dialogue systems.

In summary, our major contributions are (1) a high-quality and challenging argumentative dialogue dataset on NLP papers, (2) a new set of intent classes for interactions in argumentative dialogues, (3) a new methodology for collecting dialogues between scientists.

We believe that our contributions pave the way for building intelligent dialogue agents in scientific domains. These agents not only can foster our access to the content of scientific papers by answering our questions but also can scientifically argue about their answers.

2 Related Work

To our knowledge, there is no dataset consisting of dialogues with both exploratory (EXP) and argumentative (ARG) interactions. The most similar datasets to ours deal with QA tasks on scientific papers, such as QASPER (Dasigi et al., 2021), emrQA (Pampari et al., 2018), BioRead (Pappas et al., 2018), BioMRC (Pappas et al., 2020) and QAngaroo (Welbl et al., 2018), PubMedQA (Jin et al., 2019), and BioAsq (Tsatsaronis et al., 2015; Krallinger et al., 2020). These datasets include only the EXP questions and lack the ARG questions and answers. Moreover, questions in these datasets weakly link to each other to resemble a dialogue. In contrast, ArgSciChat consists of coherent
### Intent Definition Example

**Exploratory (EXP)**
- **Ask Info (AI)**: A question to seek information from a paper.
- **Give Info (GI)**: An answer to give information about a paper.
- **Switch Topic (ST)**: A request or proposal to shift the topic.

**Argumentative (ARG)**
- **Ask Opinion (AO)**: A question to ask about opinions.
- **Give Opinion (GO)**: An answer to provide an opinion.

---

**Table 1:** We define five intent classes for questions and answers in a scientific argumentative dialogue.

| Intent                | Definition                                           | Example                                                                 |
|-----------------------|------------------------------------------------------|----------------------------------------------------------------------|
| Exploratory (EXP)     | A question to seek information from a paper.         | What data set is used to train a classifier?                          |
|                       | A request or proposal to shift the topic.            | We consider the Partial Latin Square completion problem.              |
|                       | An answer to give information about a paper.         | Which intent would you like to know more about?                       |
| Argumentative (ARG)   | A question to ask about opinions.                    | Which do you think is the strongest point of your approach?          |
|                       | An answer to provide an opinion.                     | It is interesting that such simple techniques help.                  |

---

**3 Methodology**

Our goal is to collect synchronous dialogues about the content of scientific papers, where dialogues include EXP and ARG questions and answers.

#### 3.1 Defined Intent Set

There are no pragmatical intent sets for collecting scientific argumentative dialogues. Therefore, we define a new set of intents for interactions in such dialogues. To do so, we follow a fundamental theory of argumentative dialogues (Walton, 2008). Walton (2008) defines six different interaction types in a dialogue. By taking into account their goals, only two types are relevant to scientific dialogues: information-seeking to extract information and persuasion to discuss the claims of a scientific paper. Inspired by this theory, we define five intent classes for questions and answers that may happen in scientific argumentative dialogue. Table 1 summarizes the definitions of these intents and shows an example for each intent. Exploratory (EXP) interactions aim to seek (or provide) information. Argumentative (ARG) interactions aim to convey persuasion. It is worth noting that Walton (2008) considers other aspects of argumentative dialogues, such as “personal attacks” and “appeals to emotion”. We do not include those aspects in this work because scientific argumentation is primarily supposed to be objective.

---

**3.2 Dialogue Formulation**

In each dialogue, each subject receives either role E (expert) or P (proponent). P and E discuss the content of a scientific paper. P accesses only the paper’s title, whereas E accesses the abstract and introduction sections of the scientific paper. We limit the paper content to the abstract and introduction sections because we believe these two sections present the gist of the paper sufficiently to sustain a dialogue.
To collect EXP interactions, we instruct P to use the title to start the conversation by asking for information about the paper. E answers the question using the content of given sections of the paper. Since E’s answers should be grounded in the content of the scientific paper, we ask E to select up to two text spans from the paper that are used in generating the answer. We refer to these text spans as supporting facts. This definition resembles the ones introduced in Dasigi et al. (2021) (evidence) and Reddy et al. (2019) (rationale). We emphasize that a supporting fact does not imply any argumentative function of the text span (e.g., a fact can be a claim like $F_3$ in Figure 1). A supporting fact only helps the subjects to conduct high-quality dialogues. To encourage natural conversations, we instruct both P and E to write down their message (which could be a question or an answer) in free-form textual sentences.

A pair of a P’s and an E’s messages constitutes a dialogue turn. The E’s message and selected supportive facts are displayed to P to initiate the next dialogue turn. To collect messages with ARG intent, we encourage P and E to argue about previous dialogue turns. An explored topic of the scientific paper should be easier to understand through argumentation. Finally, to limit the occurrences of unanswerable questions and improve the coherence of the dialogues, we ask E to switch the topic of conversation using the given content of the paper.

### 3.3 Dialogue Collection

To collect scientific argumentative dialogues on scientific papers, we focus on NLP publications. While including many interdisciplinary scientific papers, this domain sets a common background between scientists as subjects. For each dialogue we ask two subjects who are experts in NLP to discuss about the content of an NLP publication.

A key challenge in implementing our data collection methodology is to optimize this process for scientists. Best-practice crowdsourcing frameworks (e.g., AMT and Upwork) lack three main features regarding dialogue collection in the scientific domain: (1) flexibility in scientific paper selection and (2) time slots scheduling, and (3) synchronous participation. To meet these requirements and incentivize scientists’ participation, we introduce an implementation for our dialogue collection methodology. We develop a web-based tool where the subjects can register and chat in a written format about an NLP publication. While subjects sign up in our tool, they confirm a consent which authorizes the usage of their dialogues and papers for any sorts of research purposes. We define the consent form by following GDPR\(^2\) and the ethics guidelines for trustworthy AI\(^3\).

To encourage subjects participate in a dialogue, we let them select a few scientific papers from an automatically retrieved list of their recent publications. This idea relieves subjects’ burden in reading the selected paper since they are the author of those papers. Moreover, from the subjects’ perspective, participating in a dialogue about their publications is an informal advertisement for their research. To align the subjects with each other time-wise, we ask subjects with role E to announce their available time slots for presenting their selected scientific paper. We then let subjects with role P choose time slots based on their time schedule and also the title of papers associated to the slots. We report the implementation details of our methodology and an example of its web interfaces in Appendix A.

We invited 31 senior and junior scientists in NLP from two large NLP groups in Europe to participate in our dialogue collection study. 23 of the invited scientists (74.2%) accepted and participated in at least one dialogue. On 20 NLP papers, we collected 68 dialogues in total. We filtered out dialogues that ended abruptly due to connection issues, resulting in 41 dialogues. We present these dialogues and their corresponding papers as the ArgSciChat dataset, which consists of 498 messages from 41 dialogues about 20 NLP papers.

### 4 Experiments

#### 4.1 Comparing with Similar Datasets

To put the ArgSciChat dataset in context, we first compare it with similar datasets. We compare dialogues in ArgSciChat with dialogues in CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), and Doc2Dial (Feng et al., 2020). These datasets include contextualized question-answering (QA) messages grounded in generic documents. Since no expert knowledge is required to explore such texts, dialogues take place between crowd contributors in these datasets. Conversely, dialogues in ArgSciChat are conducted by experts in NLP as

\(^2\)https://gdpr.eu/tag/gdpr/

\(^3\)https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
a scientific domain. Importantly, unlike the other datasets, dialogues in ArgSciChat follow EXP and ARG intents.

4.2 Analyzing Dialogues of ArgSciChat

To obtain a deep understanding of the dialogues in ArgSciChat, we perform an in-depth analysis on this dataset. In particular, we investigate the diversity of dialogues collected for a paper, the properties of selected supportive facts, and to what extent the dialogues are exploratory and argumentative.

4.3 Using ArgSciChat for Evaluating Dialogue Agents

To assess the difficulty of ArgSciChat for recent dialogue agents, we synthesize a dialogue agent to take the role of E. Consequently, we define two evaluation tasks: (i) Fact selection, where the agent selects up to two sentences from a given scientific paper as supportive facts; and (ii) Response generation, where the agent generates a free-form text to respond to a P’s message.

We evaluate LED (Beltagy et al., 2020), which is a Longformer trained to answer questions about the content of a long text. LED has been recently used for QA on scientific papers (Dasigi et al., 2021). Using our dataset, the input to LED can be: a P’s question to which we refer as a query (Q); an NLP paper (P); all interactions exchanged before Q, also known as dialogue history (H); and the supportive facts (F) selected by human subjects to generate the reference E’s answers. We thus compare the following LED configurations: LED(Q, P), LED(Q, P, H), and LED(Q, F), where terms inside the parentheses are the inputs to the LED agent.

Dasigi et al. (2021) show that LED achieves better performance when trained jointly for response generation and fact selection. Likewise, we train LED using the loss function $\mathcal{L} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{fact}}$, where $\mathcal{L}_{\text{gen}}$ is the average negative log-likelihood loss between a generated message and the reference E’s message. The $\mathcal{L}_{\text{fact}}$ loss is the average binary cross-entropy loss between the label that a neural classification layer, i.e., MLP, assigns to each sentence of the paper and the reference fact labels we defined for those sentences. To define the reference fact labels for a sentence in a scientific paper, we compare the sentence with the corresponding human-selected supportive facts. We assign a ground truth label 1 to a sentence if a supportive fact contains that sentence, otherwise 0.

We carry out a five-fold cross-validation routine for this experiment. We create fold splits such that all dialogues about the same scientific paper are in the same split. Table 2 shows the number of dialogue turns as data samples used in the training, validation, and test sets of each fold. We report the values of all hyper-parameters used to train LED in Appendix C.

For evaluation metrics, similar to Dasigi et al. (2021), for fact selection, we compute the F1 score over candidate sentences in a scientific paper against the reference supportive facts. We denote this metric as Fact-F1. For response generation, we use the token-level F1 score introduced in SQUAD (Rajpurkar et al., 2016). This metric is computed over individual words between the generated response and the reference E’s message. We denote this metric as Message-F1. Following recent suggestions for evaluating text generation systems, we also report BERTScore (BScore) (Zhang et al., 2020) and MoverScore (Mover) (Zhao et al., 2019) for the response generation task.

To put our results in context, we use a retrieval-based method as a baseline for the fact selection task. This method computes the cosine-similarity score between the vector representation of a query Q and the vector representation of each sentence in the input paper. Following Dasigi et al. (2021), we use TF-IDF with default parameters in sklearn and S-BERT to obtain the vector representations. We then rank the sentences concerning their similarity scores with the query Q and select the top-two sentences as the supportive facts. Like Dasigi et al. (2021), we estimate a statistically robust human baseline for the fact selection task with three NLP experts. In particular, each expert selects two sentences from the input paper as supportive facts for a given query Q. We would like to remark that human performance estimation process is independent of the annotation process described in Section 5.2 and involves different NLP experts to ensure sound evaluation.

| Fold   | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|--------|--------|--------|--------|--------|--------|
| Train  | 199    | 200    | 183    | 212    | 207    |
| Validation | 27    | 14     | 39     | 27     | 22     |
| Test   | 23     | 35     | 27     | 10     | 20     |

Table 2: The number of dialogue turns in each fold.
Table 3: Avg. Turns shows the average number of turns per dialogue. % MSM depicts the percentage of multi-sentence message in a dataset. Avg. Length indicates the average number of tokens per message. # Dial. indicates the number of dialogues. Supporting facts are not considered as answers.

5 Results

5.1 Comparing with Similar Datasets

Table 3 reports the characteristics of dialogues in the compared datasets. In terms of the average number of dialogue turns, dialogues in ArgSciChat are comparable with those in QuAC and Doc2Dial datasets. Compared with CoQA, dialogues in ArgSciChat contain less dialogue turns. ArgSciChat contains fewer dialogues than the other datasets. The limited number of dialogues in our dataset is symptomatic of expert domains where the data collection process is subject to tight requirements like matching scientists’ schedules. The exchanged messages in ArgSciChat are beyond text spans and consist of free-form textual sentences. In particular, ArgSciChat contains a higher percentage (50%) of multi-sentence messages than the other datasets. The limited number of dialogues in our dataset is symptomatic of expert domains where the data collection process is subject to tight requirements like matching scientists’ schedules. The exchanged messages in ArgSciChat are beyond text spans and consist of free-form textual sentences. In particular, ArgSciChat contains a higher percentage (50%) of multi-sentence messages than the other datasets. This observation is also confirmed by the average message length, i.e., the number of tokens per message.

5.2 Analyzing Dialogues of ArgSciChat

Dialogue Diversity. On average ArgSciChat contains two dialogues for each NLP paper. We study to what extent the messages (which could be questions or answers) exchanged in these dialogues are semantically diverse. To do so, we group messages in dialogues grounded in one paper into three categories: G1: semantically similar P messages; G2: E messages associated with P messages in G1; and G3: E messages and their corresponding supportive facts. For each group, we compute the semantic diversity between any pair of messages, using the cosine distance between the vector representations of messages. We obtain the vector representations of messages using S-BERT⁴ (Reimers and Gurevych, 2019). If the cosine distance between the vector representations of the messages in a pair is below a threshold, we consider the messages to be semantically diverse. We empirically found that a threshold value of 0.5 is sufficient to study dialogue diversity in ArgSciChat. Qualitatively similar results obtained with different thresholds are reported in Appendix B.

We report the average percentage of semantically diverse pairs in each group over all papers with multiple dialogues. About 90% of messages in group G1, 63% in group G2, and 43% in group G3 are semantically diverse. These results suggest that our dialogue formulation yields diverse dialogues even on identical papers. This property of ArgSciChat is important when its dialogues are used to evaluate dialogue agents (Reddy et al., 2019).

Supportive Facts Up to two supportive facts are used as groundings for E’s messages. Table 4 shows that about 61.1% (129 out of 211) of these messages are grounded in one supportive fact and 38.9% (82 out of 211) on two supportive facts. For the two-facts messages, the average sentence-based distance between the facts in the scientific paper is 5.8. This observation indicates that the supportive facts in the two-facts messages are non-adjacent sentences in the paper.

We also observe that a large portion of all the supportive facts is from the introduction section. In particular, 61% of supportive facts are from the introduction section, while 39% are from the abstract. This shows that both abstract and introduction sections of the scientific paper are explored in the dialogues.

Finally, E selects most of the supportive facts from the introduction section, denoting the exploratory intent of P. We study the distribution of supportive facts over sentences in the introduction section. We view the introduction section at the sentence level. We compute the supportive fact distribution by normalizing each sentence position.

Table 4: Studied statistics about supportive facts in ArgSciChat.
concerning the total number of sentences in the introduction section. Figure 2 shows the results. We observe that supportive facts mostly are from the middle (35%-55%) and last (70%-100%) sentences of the introduction section. The middle part of the introduction section describes the research method used in the corresponding scientific paper. The last part reports experiments and findings. We conclude that dialogues explore the scientific contributions of a scientific paper from different scientific views.

**EXP and ARG Intents** To study to what degree the dialogues are exploratory and argumentative, we instruct four NLP experts to annotate all the sentences from ArgSciChat dialogues. Specifically, each expert annotates 25% of ArgSciChat sentences. We remark that these experts are not from those that participated in our dialogue collection study. As instruction, we give them our intent definitions and their example sentences (Table 1). For the sake of the annotation budget, we compute the inter-annotator agreement on 289 sentences of 10 randomly selected dialogues. The average Fleiss’ Kappa (Fleiss, 1971) is 0.89, and the Krippendorf’s alpha (Krippendorff, 2011) is 0.83, showing that the quality of annotations is high enough to be used in our analysis of the EXP and ARG intents of dialogues in ArgSciChat.

Table 5 reports the frequency of each intent label on all the sentences expressed by P and E from all dialogues in ArgSciChat. P is more inclined to ask questions (AI), whereas E reports about facts from paper content (GI). This is compatible with our definitions for these roles, showing that our dialogue formulation was comprehensive for the expert subjects to follow. P expresses an opinion (GO) more often than E by 10%. Both P and E ask for opinion (AO) with a similar frequency (≈6%).

E switches the topic of conversation (ST) more frequently than P by about 6%, which is because E has access to more content of the scientific papers than P. Moreover, according to our formulation, E should drive the conversation to prevent any dead end. By grouping intent labels according to their corresponding intent categories (the EXP and ARG columns in Table 5), we observe that ARG intents represent nearly 23% of all the sentences. In comparison, EXP intents represent about 77% of messages. These results support the correctness of our dialogue collection methodology in achieving the goal of collecting dialogues that demonstrate EXP and ARG dialogues in a scientific domain.

**5.3 Using ArgSciChat for Evaluating Dialogue Agents**

Table 6 shows the results for both tasks. For fact selection, we observe that the performance of LED in terms of Fact-F1 is comparable with the baseline fact selection models and far from human performance on this task. This confirms the difficulty of ArgSciChat for the fact selection task. However, one reason for such a low score of LED is the small number of data samples, motivating further research in developing few-shot learning methods for training conversational agents in expert domains.

For response generation, in terms of the Message-F1 metric, LED(Q,P) achieves a 14.25% token-level F1 score, followed by the LED(Q,P,H) with a 16.14% score. LED(Q,P,H) outperforms

| Role | AI | ST | GI | GO | AO | EXP | ARG |
|------|----|----|----|----|----|-----|-----|
| P    | 63.3 | 1.2 | 7.6 | 21.9 | 6.1 | 72.0 | 28.0 |
| E    | 1.0 | 7.6 | 72.7 | 11.8 | 6.9 | 81.3 | 18.7 |
| Total | 29.5 | 4.7 | 42.9 | 16.4 | 6.5 | 77.1 | 22.9 |

Table 5: The frequency (%) of intents from each dialogue role. “Total” is the total frequency of each intent.

| Model   | Fact-F1 | Message-F1 | BScore | Mover |
|---------|---------|------------|--------|-------|
| TF-IDF  | 16.22   | -          | -      | -     |
| S-BERT  | 13.65   | -          | -      | -     |
| LED(Q,P) | 10.58  | 14.25      | 85.85  | 2.25  |
| LED(Q,P,H) | 8.50   | 16.14      | 86.00  | 4.54  |
| LED(Q,F) | -      | 19.54      | 86.64  | 8.53  |
| Human   | 51.26   | -          | -      | -     |

Table 6: LED’s scores for fact selection (Fact-F1) and response generation (Message-F1, BScore, and Mover).
Query: Sounds promising! So how well do the models perform?

LED(Q,P): We use a neural network to train the neural network to generate neural networks.
LED(Q,P,H): We use a generic, KB-agnostic annotation approach.
LED(Q,F): They use their models to improve annotation speed. What do you think about this?

Reference E: Using these in a user study, we measured an annotation time improvement of up 35%. Do you think this is useful?

Reference Facts: We show that statistical machine learning models can be used in an interactive entity linking setting to improve annotation speed by over 35%.

Table 7: A case study of the response generated by the LED agent when trained on the ArgSciChat dataset.

LED(Q,P), suggesting that prior exchanged messages in a dialogue can be beneficial in generating a response to an input query. This also shows that dialogues in our dataset are beyond a sequence of unrelated dialogue turns. LED(Q,F) outperforms LED(Q,P) and LED(Q,P,H) by at least three token-level F1 points. This score of LED(Q,F) confirms the importance of supportive facts for generating a high-quality response.

As a case study, we analyze responses LED generates for a query Q given different inputs. Table 7 shows the query, the responses generated by the examined configurations of LED, the response generated by E, and the supportive facts selected by E in a dialogue. LED(Q,F) identifies Q’s intent, which aims to ask for information about “the speed improvement of the annotation study”. The model generates a response with both EXP and ARG intents. In particular, the first sentence in the LED(Q,F)’s response has EXP intent as it provides information to answer the question. The second sentence involves ARG intents as it asks for P’s opinion, heading towards an argumentative dialogue.

6 Discussion

Our analysis of dialogues in ArgSciChat suggests that our methodology is an effective practice for collecting dialogues in expert domains. In this work, we applied our methodology to the NLP scientific domain. To demonstrate that the methodology is sound, we have run a pilot study and invited 31 NLP experts. Among the invited experts, 23 of them accepted our invitation and participated in our dialogue collection study. In future work, the methodology could be further scaled to online community events, poster sessions in online conferences, where scientists chat about their papers. The performance of LED on our dataset highlights different characteristics of ArgSciChat, such as the importance of supportive facts for response generation, the benefit of including dialogue history, and the challenge of generating messages that have both EXP and ARG intents. In particular, these characteristics motivate further research on building conversational agents in expert domains. Our dataset’s size testifies to the challenge of incentivizing scientists’ contribution to data collection. Recent large language models have shown a remarkable ability for zero- and few-shot learning (Brown et al., 2020). We believe that even a small-sized but high-quality dataset could contribute significantly to developing dialogue agents in specialized domains.

7 Conclusion

We propose ArgSciChat, the first argumentative dialogue dataset in the NLP scientific domain. In our dataset, dialogue sentences are in free-form English texts, written by NLP experts, and annotated with intent labels encoding the EXP and ARG intents of scientific argumentation. To collect this dataset, we defined a new methodology that lets scientists introduce their scientific papers, choose the topic and the time of dialogues based on their preferences. In addition to in-depth analysis of ArgSciChat, we found that ArgSciChat is a challenge for a recent dialogue agent (Beltagy et al., 2020) for the tasks of supportive fact selection and response generation. In future work, we aim at designing few-shot learning agents in the scientific domain.

Ethical Consent

We ask experts to read and confirm a consent concerning data privacy and informed consent before signing up for our tool. In the form, we explicitly
state the aim of the study and the later use of collected data. We provide detailed information to the subjects about the personal data information we require for participation and its temporary usage throughout the study. Subjects can request data deletion at any given step of the study. All subjects who agree to sign-up also consent to participate in the study.

**Limitations**

ArgSciChat is the result of a pilot study concerning 31 invited NLP experts. In particular, ArgSciChat contains dialogues about 20 scientific papers regarding a few NLP topics. Thus, dialogues in ArgSciChat are only a small sample of the set of possible dialogues grounded in scientific papers. In particular, several design choices have been considered in our data collection methodology: (a) the topic of a paper; (b) the common background of invited NLP experts; (c) the available content of a paper during a dialogue; and (d) the dialogue setting (e.g., in our implementation, dialogues had a time limit which restricted the number of interactions between subjects). We chose the NLP domain in our study since we (the authors) have expertise in this domain. This choice also facilitated the definition of a pool of NLP experts to participate in our study through our research network.

Furthermore, dialogues in ArgSciChat are grounded in scientific papers. In particular, we limit the paper’s content to the abstract and introduction sections. This choice reduces subjects’ effort to act as E, while also providing enough information to sustain a dialogue. Thus, we do not collect dialogues between subjects concerning other sections of a paper.

**References**

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuasive oNline discussions. 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), pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentai Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, June 6-11, 2021, pages 4599–4610, Online.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, November 16-20, 2020, pages 8118–8128, Online.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In Proceedings of the 54th Annual Meeting of the Association for Computational
Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, pages 4684–4690, Florence, Italy.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. 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) 2019, pages 591–602, Hong Kong, China.

Mathilde Janier and Chris Reed. 2016. Corpus resources for dispute mediation discourse. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016, European Language Resources Association (ELRA).

Qiao Jin, Bhawan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. 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) 2019, pages 2567–2577, Hong Kong, China.

Martin Krallinger, Anastasia Krithara, Anastasios Nentidis, Georgios Paliouras, and Marta Villegas. 2020. Bioasq at clef2020: Large-scale biomedical semantic indexing and question answering. In Advances in Information Retrieval, pages 550–556, Online. Springer International Publishing.

Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.

Anne Lauscher, Lily Ng, Courtney Napoleon, and Joel R. Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, pages 4563–4574, Barcelona, Spain (Online).

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. Computational Linguistics, 45(4):765–818.

Fabrizio Macagno. 2000. Types of dialogue, dialectical relevance, and textual congruity. Anthropology & Philosophy, 8(1-2):101–121.

Randall Munroe. 2013. The rise of open access. Science, 342(6154):58–59.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. A dataset of general-purpose rebuttal. 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 2019, Hong Kong, China, November 3-7, 2019, pages 5590–5600. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2357–2368, Brussels, Belgium.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan T. McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, pages 140–149, Online.

Joonsuk Park and Claire Cardie. 2018. A corpus of rulemaking user comments for measuring evaluable of arguments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Daraksha Parveen, Mohsen Mesgar, and Michael Strube. 2016. Generating coherent summaries of scientific articles using coherence patterns. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 772–783, Austin, Texas. Association for Computational Linguistics.
Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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), pages 3980–3990, Hong Kong, China.

Francesco Ronzano and Horacio Saggion. 2015. Dr. inventor framework: Extracting structured information from scientific publications. In Discovery Science - 18th International Conference, DS 2015, volume 9356 of Lecture Notes in Computer Science, pages 209–220, Banff, AB, Canada.

Allen Roush and Arvind Balaji. 2020. Debatesum: A large-scale argument mining and summarization dataset. CoRR, abs/2011.07251.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunkle, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almiritis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androustopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform., 16:138:1–138:28.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. Lang. Resour. Evaluation, 54(1):123–154.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Douglas Walton. 2008. Informal Logic: A Pragmatic Approach, 2 edition. Cambridge University Press.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. Transactions of the Association for Computational Linguistics, 6:287–302.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia. OpenReview.net.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3580–3586, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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) 2019, pages 563–578, Hong Kong, China.
A Dialogue Collection Supplementary Details

We provide details about our implementation of the described data collection methodology below. We define a time deadline for each step to ease subjects’ synchronization along the data collection process.

Sign-up We require subjects to provide the following contact information: full name and email address. We also need subjects to provide their Google Scholar profiles regarding paper selection. If no Google Scholar profile is available, subjects can still sign-up by manually submitting hyperlinks to the PDF files of papers about which they have enough expertise. The automatic paper retrieval step reports the top five subjects’ most recent papers listed in their Google Scholar profile. Subjects are then asked to select two of them to participate in the study.

Booking as E We consider time slots of one hour in length. Upon sign-up, summary information about selected papers and time slots is provided to human subjects. In particular, subjects can add the selected time slots to their calendar to ensure participation.

Booking as P Once the sign-up deadline has been met, our implementation automatically notifies subjects about the next phase via email. Additionally, a unique private authentication code is assigned and provided to each participant. The authentication code is used to identify subjects during the study while ensuring anonymity. We split each E time slot into three time slots of 20 minutes duration. Subjects are required to book four distinct time slots, each one regarding a different paper. Upon submission, summary information is provided to subjects regarding their final schedule concerning both roles.

Conducting a Dialogue Once the time deadline of the “Booking as P” step has been met, our implementation automatically notifies subjects about their final schedule. By doing so, subjects have exact information about which time slots they have to join. Our implementation also supports an automatic notification system that ensures subjects join each time slot on time.

Why Synchronous Dialogues? Given subjects’ required level of expertise and to ensure naturalness during dialogue collection, we opted for synchronous dialogues. Indeed, such a choice introduces tight requirements about subjects’ availability and corresponding attributable effort. However, asynchronous dialogues can also have potential drawbacks, such as an extended data collection period. Additionally, in an asynchronous dialogue, dialogue properties like the partial observability of P can be quickly overcome due to the absence of a time limit. Conversely, short synchronous dialogues encourage both subjects to maximize the exchanged information flow in the same way as a natural goal-oriented conversation.

Data Collection Interfaces We develop role-specific collection interfaces based on the Mephisto library. We opt for a standalone data collection application that does not impose additional participation steps apart from initial data submission. Upon connection of both subjects for a dialogue, the corresponding paper content is loaded, and dialogue roles are attributed to each subject. Coherently with the given formulation, no worker payment was considered, but an award system is devised to encourage attendance. To avoid a E message without supportive facts, the system notifies when no facts have been highlighted upon a message sent. Additionally, a hint system encourages both subjects to suggest discussing previous dialogue turns encouraging opinions exchange. Figure 3 reports the data collection interfaces for both dialogue roles. Both interfaces reflect the asymmetric nature of the dialogue by restricting information to individual roles.

B Dialogue Diversity Supplementary Details

Table 8 shows the diversity scores for each group of messages across different diversity thresholds.

| Group | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
|-------|-----|-----|-----|-----|-----|
| G1    | 0.77| 0.87| 0.92| 0.96| 0.98|
| G2    | 0.43| 0.48| 0.63| 0.73| 0.68|
| G3    | 0.18| 0.28| 0.43| 0.60| 0.81|

Table 8: Average topical diversity on dialogues on identical papers. The numbers in the header of columns represent thresholds values.

5https://github.com/facebookresearch/Mephisto
Figure 3: Data collection interfaces for (a) P and (b) E. On the right side, the title, abstract and introduction sections of the paper are reported for E. Conversely, P’s view is solely restricted to the paper title.

C Experiment Supplementary Details

In this section, we report additional details regarding our experimental setup and the described tasks.

Setup As in Dasigi et al. (2021), multiple inputs are concatenated together with a special separation token. In our experimental setup, inputs are concatenated based on the following order: (i) query (Q); (ii) dialogue history (H); (iii) scientific paper (P). We set the LED global attention mask to take into account each individual input as in (Dasigi et al., 2021). In contrast to (Dasigi et al., 2021), the lack of LaTeX source files forbids the automatic retrieval of paper section paragraphs. Instead, we work at the sentence level. The average training time is ~40 – 60 seconds per epoch depending on the given input combination. Experiments are conducted on a NVidia GeForce 2080ti 11 GB.

Model Configuration Table 9 reports the main hyper-parameters of the LED model. As in (Dasigi et al., 2021), we employ the HuggingFace model allenai/led-base-16384. Because we only use the abstract and introduction sections of a scientific paper, we re-scale the attention window size hyper-parameter according to our maximum document length. Note that the selected hyper-parameter value is the same percentage of the maximum document length as in (Dasigi et al., 2021).
| Hyperparameter          | Value          |
|-------------------------|----------------|
| Attention dropout       | 0.1            |
| Attention window size   | 700            |
| Optimizer               | Adam           |
| Learning rate           | $5 \cdot 10^{-5}$ |
| Patience                | 20             |

Table 9: The LED model hyper-parameters.