Tree-of-Debate: Multi-Persona Debate Trees Elicit Critical Thinking for Scientific Comparative Analysis

Priyanka Kargupta
UIUC
Urbana, United States
pk36@illinois.edu

Ishika Agarwal
UIUC
Urbana, United States
ishikaa2@illinois.edu

Shivam Agarwal
UIUC
Urbana, United States
shivama2@illinois.edu

Abstract

As modern technology and improved accessibility scales the depth and breadth of research exponentially, scientific discoveries have become further disconnected within and across fields. Hence, determining significant novelties, incremental findings, and equivalent approaches between works is challenging, especially when the papers are not explicitly connected through citations. In order to elicit the critical reasoning required for comprehending the contribution degree of a paper, we propose converting the papers to LLM personas which debate one another. In other words, we propose a tree-of-debate (ToD), where we focus more on the personas' comparative reasoning induced by the debate, as opposed to its final outcome. ToD can dynamically construct a debate tree to reason about fine-grained arguments discussed in scholarly articles. Through experiments on real-world scientific literature, we aim to show that Tree-of-Debate elicits informative arguments and contrasts papers, aiding researchers in their work.

CCS Concepts

• **Debate, Comparative Summarization, LLM Personas**;

Keywords

Debate, Comparative Summarization, LLM Personas

ACM Reference Format:
Priyanka Kargupta, Ishika Agarwal, and Shivam Agarwal. 2018. Tree-of-Debate: Multi-Persona Debate Trees Elicit Critical Thinking for Scientific Comparative Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Disseminating new research findings has become non-trivial with the popularity of open-access repositories. For example, arXiv received over 24,000 submissions in October 2024 [1], inundating researchers with an overwhelming volume of information. This astronomical surge in scholarly articles, coupled with the increasingly fragmented nature of research communities [2], makes it difficult to identify novel findings and discern the distinctions between related papers, especially those which lack direct citation links.

Automatically summarizing research papers has shown to help researchers in their work [9]. These methods aim to generate human-readable summaries containing only essential information from the article. State-of-the-art solutions involving large language models (LLMs) often generate high-level summaries of several articles independently. While useful, these summaries do not provide useful information for understanding the fine-grained differences between related works. Several user-studies show that users find it most helpful when they see fine-grained differences between the presented options [8, 13]. For instance, while summarizing two closely related works on pre-trained models, "BERT" [6] and "RoBERTa" [17], highlighting that the former does not use next sentence prediction, trains on ten times more data, and has better performance will help researchers in identifying the key differences.

Existing works on contrastive summarization [24] share a common pipeline: 1) construct extractive summaries of each document and 2) use the extractive summaries to identify key differences between the documents. While successful, these methods fail to reason about fine-grained claims made in full academic papers. Building on principles of Society of Mind [20], multi-persona discussions like debates present a new avenue for improving the capabilities of existing pipelines. Multi-persona debates [3, 14] mimic human like group discussions that enable more fine-grained reasoning through critical feedback and inherent comparative reasoning. However, directly applying multi-agentic frameworks to scientific comparative analysis is non-trivial. Specifically, we encounter the following unique challenges:

- **Multiple contributions.** A scientific paper often makes a contribution(s) (e.g., methodology, dataset, evaluation metric) that can be deconstructed into multiple "subideas". Some of the subideas may or may not be novel (e.g., the underlying model architecture pre-dates their work, but their fine-tuning and evaluation mechanisms are novel) and hence, each should be independently evaluated for their degree of novelty.
- **Long context challenges.** Given the lengthy nature of scholarly articles, providing a full paper in-context to a model is ineffective, as details specific to the contribution in question may get overshadowed. On the other hand, solely utilizing the paper's title and abstract lead to a high-level comparison which relies mainly on surface-level semantic dissimilarities.

In order to address the above challenges, we introduce **Tree-of-Debate** based on the following principles:

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

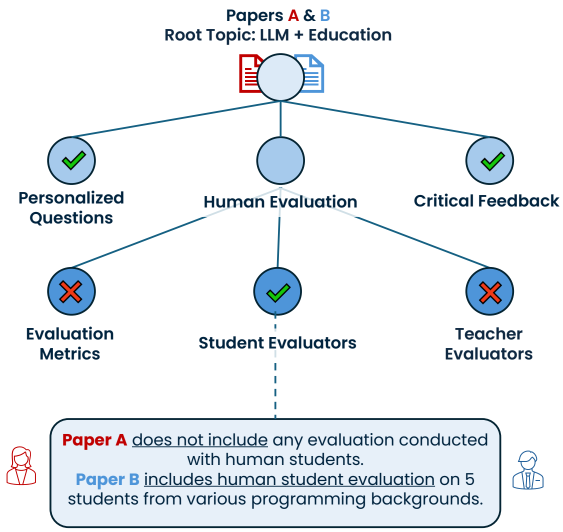

Figure 1: An example of a hierarchy based on unique and overlapping contributions of Paper A and B, specific to the root topic. A green check mark indicates that a clear "novelty winner" has been identified (e.g., Paper B includes student evaluators), whereas a red X indicates overlapping contributions between the papers (e.g., both using teacher evaluators).

(1) **Multi-persona debates elicit critical comparative reasoning.**
By converting papers into author personas which debate one another, we can incentivize a more critical analysis between the papers. Specifically, given a topic and each paper's claimed "novel" contribution towards a topic, if they now must debate whose contribution is better, then this leads to an author critically evaluating their opposition's work with respect to novelty relative to their own work.

(2) **Tree-structured debates allow for independent assessments of different contributions at varying depths.**
As mentioned above, each paper can contribute either multiple ideas or one large idea which can be partitioned into multiple subparts. Thus, it is important to evaluate the degree of novelty for each of these subparts individually, as opposed to the entirety of the paper. Hence, we can model this through a tree-structured debate, where each *node* represents a specific contribution topic being explored and debated. An *edge* between nodes indicates that there are interesting and/or unresolved points or questions brought up during the parent debate node that warrant further exploration in a child debate node. We demonstrate an example of these topical relationships in Figure 1.

(3) **Iterative retrieval throughout a debate improves fine-grained reasoning.**
Given the long-context challenges of the scientific literature domain, we motivate the need for an iterative retrieval process– specifically, where the retrieval queries are dynamically determined by the content of the debate. This way, the retrieved content can be targeted towards the specific novelty topic in question, making it easier for a persona to generate a more convincing affirmative or opposing argument. For example, as we traverse from "human evaluation" to "student evaluators" in Figure 1, the evidence pool will be updated to be more fine-grained and relevant to the subtopic.

Based on these principles, Tree-of-Debate consists of the following three steps in an effort to structure a debate between author personas, conducted by a moderator:

(1) **Self-Deliberation.** Before actually debating one another, we study the effectiveness of having each persona "prepare" for the debate: (a) each author retrieves the segments from their paper relevant to the input root topic, (b) conditioned on these segments, each author independently determines their "novel contributions" towards the root topic with supporting evidence, and (c) they retrieve any relevant segments to each other's claimed contributions from (b). Based on the author's claimed contributions, evidence, and counter-evidence, the debate moderator determines the most valuable subtopics to explore (e.g., second level of Figure 1).
(2) **Debate.** A child debate node is formed from each subtopic, where the personas debate the claim, "My paper's contributions are better for the [subtopic_name] than the opposition's". Each persona presents their initial argument on this claim, responds to one another (e.g., clarifying questions, doubts), and revises their argument based on the interaction.
(3) **Debate Node Expansion.** Based on the debate, the moderator determines if the debate node is worth expanding and exploring deeper into. Specifically, (a) was there progression in the arguments made, (b) are there any unresolved questions or doubts, and (c) was there a clear winner for novelty towards the subtopic? If so, then the node will be expanded by inputting the subtopic into (1) and continuing the same iterative process.

Our contributions can be summarized as:

- We introduce Tree-of-Debate, a general purpose multi-persona LLM debating framework, to generate fine-grained contrastive summaries.
- Tree-of-Debate can dynamically construct a debate tree to reason about fine-grained arguments discussed in scholarly articles.
- Through experiments on real-world scientific literature, we show that Tree-of-Debates elicits informative arguments and contrasts papers, aiding researchers in their work.

2 Related Works

2.1 Persona Creation & Debate

Similar to how a person's background reflects their unique abilities, recently, language models have been given personas to represent various perspectives and extract different capabilities [7]. In particular, [21] creates personas of authors based on the papers they write for *author discovery and recommendation*. Their method creates a representation of each author by using named entity recognition on their papers. By determining two authors that have the most amount of common terms between them, they match authors together for author recommendation. Our approach differs most in the task. While our approach also creates personas of each author, our persona represents only one paper per author, where [21] represents a set of papers per author. By using one paper per author to create personas, we can extract the specific novelties and incremental work between two papers. Our personas are also given a more active role as they are asked to debate and defend their paper, while the personas in [21] are only used for comparison.

There are works that create multiple personas for debate to solve reasoning problems. For example, [15] uses a persona for an affirmative and negative debater, along with a judge persona. The affirmative and negative personas talk to each other while the judge synthesizes the results of the debate into the final answer. They apply this for commonsense generation and arithmetic reasoning tasks. Unlike in self-reflection —where a model solves all the steps of a problem itself and is overly self-confident, resulting in the model's reflections to be biased towards itself— they show that with different language models producing different steps of the result, the external feedback from another model introduces more diversity and less bias in the reflection which improves task performance. While our tasks are different, we aim to further apply these results of using debate to generate better summaries of the differences between research papers. Moreover, these works use debate to reach a more accurate outcome, but in our case, the debate is the outcome.

2.2 Comparative Summarization

Generating summaries that compare two entities is challenging because of the wide variety of ways to represent differences. Most common are graph-based methods [4, 25] that use graphs to distinguish between sentences with common information and diverging information. [4] classify each of the sentences in a set of papers into one of three categories: salience, commonality, or difference[1]. They estimate scores for each sentence that reflect *salience*, *commonality*, and *difference*, and chooses sentences based on these scores. This method generates extractive summaries. While we use extractive summarization questions for the pre-debate prep, ultimately we aim to generate an *abstractive* summary that synthesizes the results of a debate.

More recent work relies on language models to extract similarities and differences. [18] fine-tunes a GPT-2 model to generate sentences that explains the difference between two scientific documents. To fine-tune, they first extract sentences —called "explanations"— with in-text citations that compare a principal document and a cited document. Next, they maximize the probability of generating the explanation given the cited document and the principal document without the explanation sentence. While this work shows good results, it only generates one sentence to compare the works and depending on the quality of the human-generated summary, it might not dive deep enough to fully capture the differences between two papers.

2.3 Related Works Generation

Multi-document summarization requires consolidating information across various sources. As the space of scientific literature becomes larger and larger, automatically going through papers to extract comparative information becomes important [4]. Specifically, [22] generates related work sections by expanding on the abstract of a given paper to create more semantically similar sentences. This essentially creates more search terms, as the extended abstract is then used as a query to retrieve relevant papers. Finally, these papers are given to a language model and, with the help of an in-context example, the related works section is generated. [27] uses a more structured approach by fine-tuning to generate good, abstractive summaries. DIR first generates summary candidates by prompting a language model with six different prompts. Next, it extracts the commonalities and differences between any candidate summary and a gold standard summary. Two separate networks are trained to maximize the commonalities and reduce the differences. These approaches, while effective, lack the flexibility that our method provides. In the first work [22], there are no guidelines on how to structure the related works. It is highly dependent on the in-context example, and if a bad example is given, the quality of the summary would suffer. Additionally, the second work [27] is dependent on the dataset, which could mean that if it needed to be adapted to a new domain of papers, the model would have to be fine-tuned again. Finally, these methods use similarity matching between the content in multiple papers [9, 16], which leads the content of the summaries to be restrictive. Alternatively, in our method, (1) the rounds of the debate act as guidelines to structure the summary, (2) our inference-time method also does not require any training, so it is domain-agnostic, and (3) we rely on the reasoning capabilities of language models to go beyond semantic similarity in an effort to identify the isomorphic property of ideas.

3 Methodology

Given two papers, we aim to determine their fine-grained scientific claims and the degree to which they are novel through a methodology inspired by formal debate. Our overall framework is presented in Figure 2.

3.1 Preliminaries

3.1.1 ***Problem Formulation.*** We are provided as input: **(a)** two papers, p_1 and p_2, and **(b)** a topic n_0 representing a theme which connects both p_1 and p_2. This can either fall under a specific dimension of scientific papers (e.g., task, methodology, experimental design; "methodologies for educational conversational agents") or a general problem (e.g., "helping students understand their mistakes").

The significance behind considering an input topic n_0 is that it provides us a starting point of the comparison– if we know that two papers already explore a specific, shared topic, we would like to focus our depth of comparison beyond this point. This will additionally help constrain the argumentation space within the debate to not discuss adjacent, but irrelevant topics. Hence, our goal is to *determine the specific novelties, incremental additions, and equivalent contributions* between two papers, p_1 and p_2, which all fall under topic n_0. These will be integrated into our output, a **debate tree** T with a corresponding **comparative summary** S. We note that while we choose to explore the comparative analysis of two papers in this work, our method is generalizable to settings beyond a 1:1 comparison (i.e., 1:n comparison).

[1](1) Salience: importance of a topic/concept in a sentence, (2) Commonality: similarity of two sentences present in two papers, (3) Difference: different of a sentence with respect to the citation text.

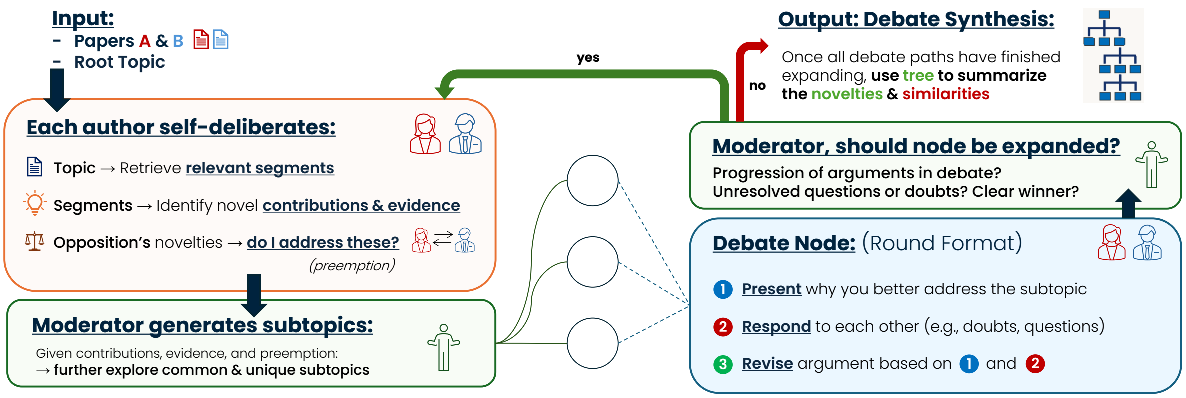

Figure 2: We propose Tree-of-Debate, a novel framework which guides a multi-persona debate using a retrieval-augmented tree. $A \rightarrow B$ within the diagram translates to the statement, "Given A, a persona arrives at B".

Each node n_i within the debate tree, T, is associated with a specific topic (this topic guides the specific debate occurring at that node). Topic n_0 serves as the root node. Furthermore, these topics can either be relevant to both p_1 and p_2 or only one of the papers (e.g., in Figure 1, only $p_i = B$ includes "Student Evaluators"). An edge from n_0 to n_1 indicates that n_1 is a subtopic of n_0 which needs further exploration to determine whether p_1 and p_2 address it.

3.1.2 ***Segment-level Retrieval***. A key underpinning to a quality debate is the individual's preparation *before* the debate and their ability to retrieve knowledge dynamically *during* the debate. Thus, we emphasize the need for a sufficient retrieval method, especially one that is adequate for scientific retrieval. We consequently utilize FlagEmbedding [30] throughout our method for our retrieval subtasks (we explored other retrieval models like SPECTER-2, but these demonstrated poorer retrieval quality). We specifically compute embeddings at the segment-level, such that the information is neither too coarse (paper-level) nor fine-grained (term-level) for proper comprehension during the debate. Hence, we pre-process the input papers with a simple segmentation method, C99 [11].

3.2 Tree-of-Debate Setup

We will conduct a multi-persona debate between two papers, p_1 and p_2. The overall debate will be conducted based on the high-level claim, *focus paper p_1 is better than the cited paper p_2 on the topic, n_i*, where each persona represents the author of a paper p. However, it is critical to note that **the final "winner" of this debate is not our focus**. This is because we should not expect a language model to be able to understand, at this point, what is a *better* or *worse* scientific claim. Rather, the primary focus of our method is the **specific reasoning** presented by the personas, *induced by the debate format*, which we argue is indicative of the progression of the debate and ultimately, the degree of novelty of a paper's claim(s). We note that we explored more novelty specific claims ("p_1's contribution towards n_i is more novel than p_2's), but this led to more surface-level semantic arguments misaligned with practical novelty.

Given that a single paper can have a breadth of claims and ideas, it is crucial that we design a structure for our debate which has the flexibility to explore these different angles independently. This will allow us to identify the presented arguments (and hence, ideas within the papers) individually and in depth. Hence, we propose a **tree-of-debate** T, which can alternatively be considered as an enriched taxonomy T (node enriched with debate + analysis, as opposed to topical nodes). Specifically, each node in the tree represents a round of debate (details provided in Section 3.2.2). The directed edge from a parent n_p to a child node n_c can be seen as a progression of the debate, deeper into one of the subtopics presented in n_p, where n_p may feature k subtopics and hence k children. Ultimately, we arrive at a leaf node once there is no progression in arguments made or there is a clear winner.

3.2.1 ***Constructing the Personas***. We use a large language model (LLM) to embody each debate persona, as it allows for more meaningful information about the paper to be integrated into its context (e.g., questions & answers about their paper, relevant extracted segments). Existing work [21] explores the construction of author personas, however it applies named entity recognition to a set of the author's papers, constructing a term-level embedding for each author. This means that the author persona is limited to term-level information, which lacks critical information about the author's paper that can be better represented at a more coarse-grained level. Furthermore, the objective of the persona is to represent the specific *paper* p_1, not the author's full collection of work. Hence, we focus more on the representation of a single paper by the author.

We summarize our personas below:

- **Authors:** An author persona should be provided with the title, abstract, and retrieved segments with respect to the given starting topic n. The retrieved segments will be updated each round based on the self-deliberation stage (Section 3.3). The focus paper's author (p_1) represents the *affirmative*: p_1 is *better* than p_2 for addressing topic n. The cited paper's author (p_2) represents the *negative*: p_2 is *better* than p_1 for addressing topic n.
- **Moderator:** Using the same underlying model as the author personas, the moderator has the following roles: (1) determine subtopics that would be important to debate for determining the papers' similarities and differences, (2) judge

the progression of the debate based on the substance of the authors' arguments, and (3) synthesize the debate into a comparative summary.

3.2.2 ***Tree Format.*** Given the root node topic, n_0, each author deliberates on their respective novel contributions towards n_0 (**self-deliberation**). From this, the moderator determines the first level of k subtopics which the authors should debate on, forming k debate children. Each debate child $\in T$ (descendant of n_0) contains three turns of inter-persona interactions. The overall format for node $n_{i\neq 0}$ is detailed below:

(1) **Debate**: Each persona p_a takes a turn presenting their argument (p_a is better than p_b for topic n_i) and their reasoning (prompt in Table 9). Then, each persona will have a turn to respond to the opposing side's argument (e.g., ask clarifying questions, mention doubts, address any of the opposition's counter-arguments; prompt in Table 10). Finally, each persona must state their revised argument based on the opposition's points (prompt in Table 11).
(2) **Determine Expansion** ***(Section 3.4)***: The moderator must determine whether the debate indicated a progression of arguments or featured unresolved questions/doubts and, hence, would benefit from a follow-up round of debate.
(3) **Self-Deliberation (*Section 3.3*)**: If the moderator determines that the debate node should be expanded, each persona p_a will retrieve the segments S_a^i from their paper that are relevant to the topic n_i. Then, each will generate a set of k primary claims C_a^i for their paper's novel contributions towards n_i, cite any corresponding evidence $E_a^i \subseteq S_a^i$, and undergo a preemptive collection of counter-evidence $\widetilde{E}_a^i$. Given C^i, E^i, and $\widetilde{E}^i$, the moderator will then determine a set of k subtopics to explore in the next level of children.

3.3 Self-Deliberation

Self-deliberation has been used as an argumentative, rhetorical strategy [26]. It allows an individual to "argue with oneself" and consider alternative perspectives in order to arrive at the best conclusion with substantive justifications. We apply this strategy to a multi-persona debate setting, where we argue that in order for LLM-based personas to have a meaningful debate on fine-grained topics, self-deliberation is essential. Given topic node n_i and paper $p_{a\in\{1,2\}}$, this involves:

(1) **Retrieving relevant segments,** S_a^i from p_a based on proximity to n_i. This is helpful for identifying more fine-grained, relevant information specific to n_i.
(2) Based on the topic n_i and S_a^i, determining a set of k claims $c_j \in C_a^i$ on the **novel contributions** that p_a makes towards n_i. For each contribution c_j, persona p_a provides a title, description, and a list of mapped evidence $E^i_{(a,j)} \in S_a^i$. We provide the prompt in Table 7.
(3) In real-world debate preparation, typically, an individual would consider the oppositions' claims in order to strengthen their own (**preemption**). Similarly, we choose to expose p_b's contributions to p_a, such that p_a can undergo an additional round of gathering evidence $\widetilde{E}_a^i$, in hopes of identifying more targeted information in their own papers.
(4) Based on $E^i, \widetilde{E}^i$, and C_i, the moderator generates a list of **subtopics** that should be further explored.

Retrieving relevant segments. For each paper $p_{a\in\{1,2\}}$, we first perform retrieval conditioned on the node topic n_i. For example, given two papers and the root topic, "LLM + Education", we initially retrieve the top δ segments $S^i_{a\in 1,2}$ relevant to this query topic n_i. As mentioned in Section 3.1.2, we utilize FlagEmbedding [30] in order to rank the segments and consider the top δ of them. This provides the initial pool of "evidence", from which each paper will form their preliminary novelty claims.

Preemption. For each $c^i_{(b,j)} \in C^i_b$ (the set of p_b's claimed novel contributions) and corresponding evidence E^i_b, we first retrieve any relevant segments $\widetilde{E}^i_{(a,j)}$ within p_a's paper. We consider each contribution $c^i_{(b,j)}$ as a separate retrieval query, concatenating its title and description. Furthermore, in order to ensure that each $e \in \widetilde{E}^i_{(a,j)}$ is truly relevant to the opposition's claim $c^i_{(b,j)}$, we pass it through an LLM-based filtration step. Specifically, we prompt the LLM to evaluate the following conditions (either outputting "yes" or "no" for each):

(1) **Does e support p_b's claim?**
(2) **Does e refute p_b's claim?**
(3) **Does e clarify p_b's claim?**
(4) **Is e irrelevant to p_b's claim?**

While redundant, we notice that explicitly including (4) as an option helps with filtration performance. Overall, if either (1-3) are true or (4) is false, then e will be filtered out. If $|\widetilde{E}^i_{(a,j)}| = 0$, then we will indicate that p_a does not address p_b's claim, $c^i_{(b,j)}$. We provide the prompt in Table 8.

Subtopic Generation. Given each paper's title and abstract, C_i, E^i, and $\widetilde{E}^i$, the moderator generates a list of k **subtopics** $n \in N^i_{\text{children}}$ that should be further explored. The moderator is also simultaneously asked to map at least one contribution c^i (from either p_1 or p_2's set of novelty claims) to each subtopic. In other words, each subtopic of n_i forms a child debate node that aims to explore either an overlapping topic between p_1 and p_2 (e.g., "Human Evaluation" in Figure 1) or a seemingly unique topic to one of the papers (e.g., "Personalized Questions" in Figure 1) which should be further confirmed as a unique novelty. The expansion prompt is provided in Table 4 of the Appendix.

3.4 Debate Tree Expansion & Synthesis

We re-iterate that, while we motivate the personas to examine whether their paper proposes a *better* idea than their opposition's within the debate, this is merely to incentivize the personas to: (1) focus more on the ***reasoning behind the idea*** proposed, and (2) induce debate regarding the ***novelty behind the ideas***, relative to each other. In other words, we hypothesize that two very similar ideas (e.g., "Human Evaluation" in Figure 1) will have more back-and-forth– a *longer* debate on whether one is better than the other. Conversely, a novel approach proposed, or new problem tackled in p_1, relative to p_2, will potentially have a *shorter* debate as ideally, the moderator will determine that p_2 does not have a strong argument as it does not address the p_1's claim (e.g., "Personalized Questions" in Figure 1). In order to facilitate and synthesis the tree expansion process, we detail the moderator's core tasks below:

Determining Round Depth Expansion. Given the arguments made in the current round of the debate n_i, the moderator should determine the following and provide an explanation:

(1) **progression_of_arguments**: Are these arguments sufficiently different enough to necessitate further debate? Are there new, deeper concepts being discussed between the two sets of arguments?
(2) **meaningful_questions**: Within the debate history, each author acknowledges each other's arguments and may ask clarifying questions accordingly. Do you believe that the clarifying questions have not been sufficiently addressed already and would be important to answer through further debate? If there are no questions raised in the debate history by either author, return False.
(3) **clear_winner**: Do you believe that it is clear that one author has won the debate, and it does not need to be further deconstructed (in order to determine which components within each author's contributions are truly better)?

If **(1)** or **(2)** are true, or **(3)** there is not a clear winner, then the moderator will continue to the self-deliberation stage to eventually discover new subtopics and expand n_i. If not, the moderator stops further expanding this debate path beyond n_i. We will also set a maximum tree depth that is sufficiently large for a quality debate (depth = 3). We provide our full prompt for this in Table 5 of the Appendix.

Debate Synthesis. Once the Tree-of-Debate (ToD) has converged (all debate paths have been adequately expanded), the moderator utilizes the overall debate tree in a structured format to synthesize the debate into a paragraph-long comparative summary.

The debate tree is provided in-context, with each node in the tree formatted as the following:

- **Node Topic Title**: <this is the moderator-generated subtopic title (if not the root node)>
- **Node Topic Description**: <this is the moderator-generated subtopic description (if not the root node)>
- **Author p_1's Revised Argument**: <this is p_1's revised argument at the end of the debate round>
- **Author p_2's Revised Argument**: <this is p_2's revised argument at the end of the debate round>

We specify that the debate synthesis should feature the similarities and differences between the papers according to the debate tree. The output summary should be structured with the papers' novelty similarities explained first and their novelty differences afterward–with more emphasis on the latter. In the Appendix, the prompt is included within Table 6 and an example of a subtree is provided in Appendix B.

4 Experimental Design

We use `nvidia/Llama-3.1-Nemotron-70B-Instruct-HF` as our base model for all experiments. We use vLLM [12] and Outlines [29] for model serving and constrained decoding. We sample from the top 1% of the tokens and set the temperature between $(0, 0.5)$ to trade-off between deterministic and creative generation. We select the best hyperparameter based on the test set results for all methods. We set $\delta = 5$, $k = 3$, and the maximum depth is 3.

4.1 Datasets

We collected a set of 14 pairs of CS NLP papers, upon which we perform human evaluation. While selecting papers, we chose papers that we are highly familiar with, so that we could perform a detailed and informed human evaluation. In addition, the papers can differ in (1) methodology (applied to the same task) or task (applied to similar motivations) and (2) explicitly have a citation link between them or not. Of the papers with a citation link, five differ in methodology and four in task; of those without a citation link, three differ in methodology and two in task.

4.2 Baselines

We compare Tree-of-Debate (ToD) with the following prompting-based baseline methods. We use the same base language model for all comparisons.

- **Single-stage**: We prompt an LLM with the title, abstract and introduction sections of both focus and opposition papers. We prompt the model to directly generate a contrastive summary of the two papers [19].
- **Two-stage**: We first instruct an LLM to individually summarize each paper based on the title, abstract and introductions. We then use the generated summaries to prompt the model to generate a contrastive summary [31].

To contextualize improvements from each component in Tree-of-Debate we construct the following ablative methods:

- **ToD (No Tree)**: We remove the tree structure from Tree-of-Debate by merging child arguments into one. We do so by concatenating the topics and descriptions of the child subtopics and tag them to distinguish the topics. In each debate round, the model is prompted with the combined subtopic and its corresponding description.
- **ToD (No SD)**: We remove self-deliberation (SD) to test the impact of iterative retrieval based on debate progression. We do so by prompting the model with title, abstract, and introduction of each paper instead of retrieving based on the subtopic.

4.3 Metrics

We use human evaluation to assess the qualities of the summaries. Our metrics are as follows:

- **Factuality**: *How factual is the summary?* Each sentence is given a 1/0 binary score for factual or not, and the scores are averaged across the summary.
- **Depth**: *Does the summary go deeper than the topic?* This is evaluated using the following Likert scale:
  (1) No, the topic is not covered at all.
  (2) No, more vague topics are covered.
  (3) No, only the given topic is covered.
  (4) Yes, but the topics discussed are not deep/granular enough.
  (5) Yes, the topics discussed are deep/granular enough.
- **Completeness**: *Does the summary seem comprehensive and complete?* This is evaluated using the following Likert scale:

(1) No, the summary misses (MULTIPLE) major points.
(2) No, the summary misses a (SINGULAR) major point.
(3) Somewhat, the summary misses minor points.
(4) Yes, the summary covers the major points.

4.4 Results

Performance Comparison. Table 1 shows the performance of Tree-of-Debate compared with various baselines over factuality, depth of debate and completeness. We observe that users found Tree-of-Debate summaries more factual, complete and granular compared to prompting based methods. This observation indicates that, multi-persona debates help carefully analyze pairs of papers and uncover more fine-grained arguments.

Next, we probe the impact of each component in Tree-of-Debate to contextualize the improvements. We note that the tree-structured debates help the most to uncover critical comparative reasoning over the papers. Tree-structured debates allow Tree-of-Debate to disentangle various contributions of the papers and critically explore subparts. We also observe that iterative retrieval and self-reasoning (self-deliberation) is helpful to uncover important contributions and their evidence to better participate in a specific subtopic. Papers that have an explicit citation link usually have a larger content overlap [5, 23], which requires deeper debates to uncover subtle differences. Through the tree structured multi-persona debates, Tree-of-Debate is able to critically reason and explore several subtopics to elicit fine-grained comparative reasoning.

We note that ToD (No SD) features a significantly highly completeness score than all other methods. This is likely due to it being provided the title, abstract, and introduction of each paper (similar to the baselines), as opposed to solely the title and abstract, like ToD and ToD (No Tree). The introduction often includes all the domain topics one would expect from the comparative summary (high recall, but potentially lower precision with respect to the round topic). On the other hand, ToD and ToD (No Tree) rely on the retrieval model identifying top δ segments relevant to the specific round topic, leading to potentially a high dominant topic precision, but lower recall, thereby reducing the completeness score.

Impact of citing vs non-citing pairs. Table 2 measures the impact on the depth of a debate over papers with an explicit citation link (citing papers) and without a link (non-citing papers). We observe that on average, methodology-based comparisons lead to deeper debates. This observation is likely because machine learning articles have a larger emphasis on components like architecture, leading to more deeper contrasts between pairs. On the other hand, task based differences can be identified at a high-level without going significantly deep into the paper. We also note that Tree-of-Debate explores deeper subtopics between citing papers as compared to non-citing papers but overall, its performance is much more consistent compared to the baselines and ablation without the tree. This motivates the robustness of a tree-based approach.

5 Case Studies

In this section, we qualitatively evaluate Tree-of-Debate. We draw summaries from one example, but see that our findings are consistent across other data samples in our dataset. The comparative summary pair we use consists two papers: TreeInstruct [10] and BRIDGE [28]. Both papers develop LLM-based instructors that guide students towards the answers to their questions, rather than give away the answers. Our given topic is "Helping Students Fix Their Mistakes".

Method	Factuality	Depth	Completeness
Single-Stage	85.43	83.92	72.71
Two-Stage	84.51	89.08	84.79
ToD (No Tree)	87.19	79.33	69.58
ToD (No SD)	90.09†	94.83†	**94.69**
ToD	**92.50**	**96.25**	88.43†

Table 1: Quality of the summaries for the baselines and ablation. Bold indicates the best performance, and † indicates second-best.

	Methodology		Task	
Method	Cites	No cite	Cites	No cite
Single-Stage	84.00	86.67	85.00	80.00
Two-Stage	88.00	93.33	85.00	90.00
ToD (No Tree)	64.00	93.33	80.00	80.00
ToD (No SD)	100.00	93.33	100.00	90.00
ToD	100.00	100.00	95.00	90.00

Table 2: Impact on depth of debate over papers citing each other and not split across methodology and task based ML papers.

5.1 Argument Evolution

The below tree outlines the topics discussed during the debate between TreeInstruct and BRIDGE personas. It contains keywords from the papers, which we would correctly expect to be present in the debate: "Socratic Questioning", "Planning", "Error Remediation", "Domain-Specific", "Personalized Learning", "Adaptive Feedback", and etc.

```
Helping Students Fix Their Mistakes
 ├── Guided Error Resolution via Socratic Questioning
 ├── Addressing Interdependent, Complex Errors
 │    ├── Addressing Interdependent Errors via Planning
 │    ├── Novel Datasets for Error Remediation
 │    │    ├── Dataset Difficulty and Realism in Error Remediation
 │    │    ├── Domain-Specific Error Remediation Datasets
 │    │    └── Incorporating Expert Decision-Making into
 │    │        Error Remediation Datasets
 │    └── Expertise in Educational Guidance
 └── Personalized Learning through Adaptive Feedback
      ├── Adaptive Feedback in Personalized Learning
      ├── LLM Optimization for Educational Domains
      └── Addressing Knowledge Gaps in Math Tutoring
```

Recalling Figure 1, our assumptions regarding the tree structure were that: (1) more **similar** contributions would be discussed at **deeper** levels in the tree, and (2) more **novel/different** contributions would appear at **higher** levels. These assumptions are confirmed in the above directory tree.

One of TreeInstruct's most novel contributions is that it uses Socratic questioning[2]– hence, this leaf node remains at a shallow depth. The most incremental contribution between the papers is the datasets: while they cover different domains (code debugging and math), they are both collected from expert human sources and are focused for error resolution via instruction. As expected, this leaf node is present at the maximum tree depth.

As a middle ground, both papers' approach to "Personalized Learning through Adaptive Feedback" are slightly differing. BRIDGE () uses pre-defined categories that describe different mistakes that students tend to make in math problems. On the other hand, TreeInstruct () provides the language model with the freedom to determine student errors through self-reflection. Finally, TreeInstruct is a multi-turn conversational instructor, while BRIDGE is a single-turn instructor. These ideas are reflected within medium-length paths.

To develop the argument regarding "Personalized Learning through Adaptive Feedback", the Moderator () first asks the personas to discuss the use of adaptive feedback:

> : "Discuss the role of adaptive feedback in enhancing personalized learning experiences..."
> : "TreeInstruct provides superior adaptive feedback through its **multi-turn planning** algorithm, dynamically constructing question trees based on student responses and knowledge states..."
> : "Bridge utilizes **expert decision-making models** to provide adaptive feedback... our method directly addresses the novice-expert knowledge gap, enabling effective remediation strategies..."

At this point, the Moderator is not satisfied with the depth of arguments and asks them to expand:

> : "Debate the effectiveness of adaptive feedback in enhancing personalized learning experiences..."
> : "Our approach dynamically constructs a Socratic question tree, tailoring the educational experience by **personalizing the sequence and type of questions posed** to learners... the opposition's reliance on expert decision-making processes, although effective, **may not be as scalable** or adaptable to diverse learning needs..."
> : "Our paper demonstrates that incorporating expert decision-making processes into LLMs significantly improves the quality of adaptive feedback..."
> : "...our approach's focus on dynamic Socratic questioning effectively addresses the quality of feedback through its **iterative, student-centric process**... Can the opposition **clarify how their expert-driven approach ensures similar student-centricity** and skill development?"
> : "...our expert-driven approach inherently promotes these aspects by **guiding LLMs to support the student's problem-solving process** rather than superficially engaging with their final answer... Can the opposition provide evidence on how their dynamic **Socratic questioning approach measures the effectiveness of fostering** these skills in students?"
> : "Our approach dynamically constructs a **Socratic question tree** to [not only] **personalize the educational experience** but also effectively **[foster] critical thinking** and problem-solving skills..."
> : "Our approach uniquely integrates expert decision-making processes into LLMs to... **foster critical thinking and problem-solving skills** through its focus on supporting the student's problem-solving process."

At this point in the debate, both personas have addressed how their approaches provide adaptive feedback to students for personalized learning. They have argued how their respective methods, while differing, reach the exact same goal. This is where the debate path ends. Throughout this debate, the personas presented their own arguments regarding the topic, were able to ask clarifying questions to the other persona, and present new arguments that not only strengthen their own claims but tailor them towards the claims/arguments of the other persona.

5.2 Baselines vs Tree-of-Debate

Table 3 contains summaries from the baseline, ablations, and our final method. Below, we qualitatively compare each summary (comparing TreeInstruct (Author 0) and BRIDGE (Author 1)), pointing out the weaknesses and strengths, and show how our method is able to address all the issues brought up in the baseline summaries.

The top left contains the Two-Stage baseline. The Two-Stage baseline tends to contain near-copy phrases from the paper, resulting in **an overly specific, extractive and unnatural summary** (an example is the first line highlighted in red: "Methodological Approach: Author 0 employs..."). As a result, the differences that are extracted are not explained very well, requiring more work to understand the terminology-heavy summary. It also makes vague claims near the end of summaries (example is the second line highlighted in red: "Despite overlapping in their pursuit of enhancing educational experiences..."). **The overall structure results in a suboptimal summary**.

Next, the top right box contains the summary for ToD (No Tree). The use of the debate format improves the quality of the generated claims. Unlike in the Two-Stage summary, it does not contain many extractive phrases, however **the structure of the debate is still fine-grained to coarse-grained**. Intuitively, the summaries should develop coarse-grained claims into fine-grained arguments. Moreover, there are **slight hallucinations** (examples are in the second and third lines highlighted in red: "Bridge on continuous expert..." and "TreeInstruct is designed..."). Still, the conclusion (last sentence) of the summary is not as vague as the conclusion from Two-Stage, but it still does not capture the intricacies of the two methods well enough.

Subsequently, the summary for ToD (No SD) is on the bottom left. The benefits of the tree are drastic, as the summary starts by

[2] BRIDGE does not always use Socratic questioning

Two-Stage	ToD (No Tree)
Both papers focus on enhancing educational experiences through innovative technological approaches. Both utilize novel methodologies to address specific learning challenges (code debugging for Author 0 and math remediation for Author 1). Both papers emphasize the importance of real-world validation and case studies to assess the effectiveness of their proposed solutions. Topic Focus: Author 0 focuses on code debugging with TreeInstruct, while Author 1 addresses math mistake remediation with the Bridge method. Methodological Approach: Author 0 employs state space estimation and dynamic tree-based questioning, whereas Author 1 utilizes cognitive task analysis and decision-making models. Evaluation Metrics: Author 0 measures performance on debugging datasets and student engagement, whereas Author 1 assesses the preference for LLM responses with expert decisions and conducts lexical analysis. Target Audience: Author 0's work is geared towards students of varying skill levels in code debugging, whereas Author 1 focuses on 1st-5th grade students in Title I schools for math tutoring. While both papers contribute to the educational technology landscape, Author 0's TreeInstruct offers a novel approach to code debugging through adaptive Socratic questioning, and Author 1's Bridge method innovatively embeds expert thought processes into LLMs for math remediation. Despite overlapping in their pursuit of enhancing educational experiences, the papers diverge significantly in their topics, methodologies, and target audiences, indicating a broad and diverse landscape in educational technology research.	Papers on adaptive guidance systems debate similarities and differences for effective Socratic questioning. While both papers aim to enhance adaptive guidance, they differ in approach: TreeInstruct (Author 0) emphasizes adaptive, expert-agnostic planning, leveraging dynamic question trees and state space-based planning to address interdependent errors, whereas Bridge (Author 1) focuses on expert-informed decision-making, incorporating domain-specific expert knowledge for more effective Socratic questioning. A key difference lies in their handling of unforeseen errors, with TreeInstruct relying on dynamic adaptation and Bridge on continuous expert annotation and model fine-tuning. Another distinction is their applicability: TreeInstruct is designed for diverse educational domains, while Bridge initially focused on math education, later expanding to other domains. Both authors eventually converge on a hybrid approach, combining expert guidance with adaptive planning for a more comprehensive solution. Despite this, their initial differences in methodology and error handling strategies remain a focal point of the debate. The debate highlights the importance of both adaptive planning and expert guidance in effective adaptive guidance systems. Ultimately, the synthesis of both approaches offers a more robust solution for personalizing student learning experiences.
ToD (No SD)	**ToD (Final)**
The papers share similarities in their focus on adaptive, student-centered approaches for mistake correction, leveraging AI-driven tools to enhance personalized learning experiences. Both papers emphasize the importance of context-aware feedback and expert guidance in promoting deeper understanding. However, the papers diverge in their novelties. Author 0's paper, 'Instruct, Not Assist,' uniquely integrates adaptive, multi-turn Socratic questioning with state space estimation, providing a more holistic and adaptable learning experience. In contrast, Author 1's paper, 'Bridging the Novice-Expert Gap,' focuses on modular, expert-guided decision-making, offering a more explicit and scalable approach to addressing complex, interdependent mistakes. While Author 0's approach excels in handling ambiguous student responses, Author 1's method provides more nuanced, human-centric guidance. The papers also differ in their evaluation metrics, with Author 0 emphasizing robust state space estimation and Author 1 highlighting the importance of contextual relevance. Ultimately, the papers present complementary approaches, with Author 0's work prioritizing adaptability and Author 1's work emphasizing expert-driven contextualization.	The papers by Author 0 and Author 1 share similarities in their focus on adaptive error resolution strategies, emphasizing the importance of context-sensitive decision-making and expert knowledge in enhancing student learning outcomes. Both papers also highlight the limitations of current LLMs in educational settings and propose novel methodologies to address these challenges. However, the papers diverge in their approaches to achieving these goals. Author 0's paper, 'Instruct, Not Assist,' uniquely integrates dynamic planning and hierarchical questioning to provide a more nuanced understanding of student thought processes, scaling to accommodate diverse student populations. In contrast, Author 1's paper, 'Bridging the Novice-Expert Gap,' leverages expert decision-making models to inform adaptive error resolution strategies, offering a more domain-agnostic framework. The papers also differ in their evaluation frameworks, with Author 0's approach utilizing a state space-based planning algorithm and Author 1's method incorporating transfer learning and domain-specific expert annotations. Furthermore, Author 0 emphasizes the importance of adaptive Socratic questioning, while Author 1 highlights the value of expert-guided decision-making in enhancing educational support. Ultimately, the papers present distinct novelties in addressing the novice-expert gap, with Author 0 focusing on adaptive structured planning and Author 1 on context-aware expert decision embedding.

Table 3: Summaries from four methods: Tree-of-Debate, our ablations, and the Two-Stage baseline. The papers being compared are TreeInstruct and BRIDGE: they develop LLM-based instructors that guide students towards the answers to their questions rather than give away the answers.

discussing the high-level summaries, and breaks down the individual fine-grained differences. This is much less extractive and more abstractive. **Using the tree structure along with the debate allows each argument to be explored further**– this is evident as after each claim, an explanation of why it matters follows (example is the line highlighted in green: "While Author 0's approach excels in..."). Still, a few of these explanations are vague and **do not reveal the true underlying motivation of the claims** (highlighted in red).

Finally, the summary for ToD (our final method) is in the bottom right box. With the self-deliberation, it was able to extract a short phrase of the motivation behind both works (the "limitations of current LLMs in educational settings"). The arguments are developed from **high-level claims to low-level**, technical concepts. The **facts are correctly identified** and do not contain any hallucinations. Moreover, the explanations preceding the claims also **reveal the underlying motivation** behind the specific novelty. Finally, the concluding sentence explains the exact difference between the two works.

6 Limitations & Future Work

We explore some limitations of our work. Within each debate round (each persona presenting their arguments, responding to one another, and revising their arguments), we note that a crucial element to a productive debate round is each persona providing meaningful feedback (e.g., doubts, clarifying questions) within the "respond" stage. However, the quality of this critical response may vary based on the difficulty of the task (e.g., a more fine-grained topic that has no presence within the model's existing pre-training dataset) and/or the size of the model.

Furthermore, our current evaluation metrics may not be sufficiently comprehensive. As mentioned in the case study, Single-Stage and Two-Stage baseline summaries were extractive and contained short phrases from the paper abstract/introduction. Technically, these summaries were correct, but oddly structured and unnatural. Hence, they received high scores during the evaluation, even though we saw that Tree-of-Debate summaries were far better in quality. To fix this, we propose to introduce preference-based metrics in our evaluation.

We will be extending Tree-of-Debate to other general, complex reasoning tasks which can exploit our tree-based decomposition and debate-based critical feedback. For instance, complex quantitative reasoning problems can often be decomposed into several sub-problems, which can be represented within our tree structure. Each persona can instead represent a different approach to solve that specific sub-problem. We can also consider extending this to a negotiation setting, where there are various aspects to consider when determining the optimal compromise. For example, two parties can negotiate on a price of a car with respect to its make, model, mileage, etc. Each of these can be explored within their own respective subtrees.

7 Conclusion

Automatic summarization is crucial for the vast amount of research being published. To combat this, we propose Tree-of-Debate: a structured methodology to comparatively summarize two bodies of work by representing paper authors as personas. By using personas, we can use a debate format to extract the key differences and similarities between papers through conversation. Our methodology, while evaluated on the setting of comparing a pair of singular papers, is generalizable to a set of papers. We show how breaking down the debate into a tree of debate rounds (each round is a debate round, and children nodes are rounds that discuss finer-grained details) yields abstractive, well-organized summaries that have adequate claim depth and contain fewer hallucinations. We use human evaluation for quantitatively measuring the quality of summaries and present qualitative case studies to back up our contributions and show that Tree-of-Debate significantly outperforms baselines.

References

[1] arXiv. 2024. arXiv Monthly Submission Statistics. https://arxiv.org/stats/monthly_submissions. Accessed: 2024-10-13.

[2] Stefano Balietti, Michael Mäs, and Dirk Helbing. 2015. On disciplinary fragmentation and scientific progress. *PloS one* 10, 3 (2015), e0118747.

[3] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).

[4] Jingqiang Chen, Chaoxiang Cai, Xiaorui Jiang, and Kejia Chen. 2022. Comparative Graph-based Summarization of Scientific Papers Guided by Comparative Citations. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 5978–5988. https://aclanthology.org/2022.coling-1.522

[5] Blaise Cronin. 1981. The need for a theory of citing. *Journal of documentation* 37, 1 (1981), 16–24.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142* (2023).

[8] Beliz Gunel, James B Wendt, Jing Xie, Yichao Zhou, Nguyen Vo, Zachary Fisher, and Sandeep Tata. 2024. STRUM-LLM: Attributed and Structured Contrastive Summarization. *arXiv preprint arXiv:2403.19710* (2024).

[9] Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2023. What's New? Summarizing Contributions in Scientific Literature. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 1019–1031. https://doi.org/10.18653/v1/2023.eacl-main.72

[10] Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani-Tur, and Jiawei Han. 2024. Instruct, Not Assist: LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging. arXiv:2406.11709 [cs.CL] https://arxiv.org/abs/2406.11709

[11] Sukhpreet Kaur and Kamaljeet Kaur Mangat. 2013. Comparative Analysis of C99 and TopicTiling Text Segmentation Algorithms. *International Journal of Research in Engineering and Technology (IJRET)* 2, 9 (2013), 529–534.

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles.*

[13] Kevin Lerman and Ryan McDonald. 2009. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers*. 113–116.

[14] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118* (2023).

[15] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118* (2023).
[16] Meng-Huan Liu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Contributionsum: Generating disentangled contributions for scientific papers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 5351–5355.
[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019). https://api.semanticscholar.org/CorpusID:198953378
[18] Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining Relationships Between Scientific Documents. arXiv:2002.00317 [cs.CL] https://arxiv.org/abs/2002.00317
[19] Anna Martin-Boyle, Aahan Tyagi, Marti A. Hearst, and Dongyeop Kang. 2024. Shallow Synthesis of Knowledge in GPT-Generated Texts: A Case Study in Automatic Related Work Composition. *ArXiv* abs/2402.12255 (2024). https://api.semanticscholar.org/CorpusID:267751245
[20] Marvin Minsky. 1988. *Society of mind.* Simon and Schuster.
[21] Jason Portenoy, Marissa Radensky, Jevin West, Eric Horvitz, Daniel Weld, and Tom Hope. 2022. Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery. arXiv:2108.05669 [cs.DL] https://arxiv.org/abs/2108.05669
[22] Zhengliang Shi, Shen Gao, Zhen Zhang, Xiuying Chen, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2023. Towards a Unified Framework for Reference Retrieval and Related Work Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5785–5799. https://doi.org/10.18653/v1/2023.findings-emnlp.385
[23] Mikhail V. Simkin and Vwani P. Roychowdhury. 2005. A mathematical theory of citing. *J. Assoc. Inf. Sci. Technol.* 58 (2005), 1661–1673. https://api.semanticscholar.org/CorpusID:17136642
[24] Thomas Ströhle, Ricardo Campos, and Adam Jatowt. 2023. Contrastive text summarization: a survey. *International Journal of Data Science and Analytics* 18, 4 (Aug. 2023), 353–367. https://doi.org/10.1007/s41060-023-00434-4
[25] Thomas Ströhle, Ricardo Campos, and Adam Jatowt. 2023. Contrastive text summarization: a survey. *International Journal of Data Science and Analytics* (2023), 1–15.
[26] Christopher W Tindale. 2020. Self-deliberation and the Strategy of the Pseudo-dialogue. *Co-herencia* 17, 32 (2020), 159–178.
[27] Pancheng Wang, Shasha Li, Dong Li, Kehan Long, Jintao Tang, and Ting Wang. 2024. Disentangling Instructive Information from Ranked Multiple Candidates for Multi-Document Scientific Summarization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2028–2037.
[28] Rose E. Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes. arXiv:2310.10648 [cs.CL] https://arxiv.org/abs/2310.10648
[29] Brandon T Willard and Rémi Louf. 2023. Efficient Guided Generation for LLMs. *arXiv preprint arXiv:2307.09702* (2023).
[30] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]
[31] Ming Zhang, Jiyu Lu, Jiahao Yang, Jun Zhou, Meilin Wan, and Xuejun Zhang. 2024. From coarse to fine: Enhancing multi-document summarization with multi-granularity relationship-based extractor. *Inf. Process. Manag.* 61 (2024), 103696. https://api.semanticscholar.org/CorpusID:268175720

A Prompts

In this appendix, we provide each of the prompts.

You are a fair and balanced moderator of a debate between two authors determining their respective novel contributions towards the following topic:
Topic: `<topic>`
Topic Description: `<topic description>`

Here are the two papers and their claimed novel contributions with corresponding evidence:

Author 0 Paper Title: `<Author 0 Paper Title>`
Author 0 Paper Abstract: `<Author 0 Paper Abstract>`
Author 0 Paper's Contribution #1: `<contribution statement>`: `<contribution topic>`
Author 0 Paper's Contribution #1 Evidence: `<contribution evidence>`
... (more evidence and contributions)
Author 1 Paper Title: `<Author 0 Paper Title>`
Author 1 Paper Abstract: `<Author 0 Paper Abstract>`
Author 1 Paper's Contribution #1: `<contribution statement>`: `<contribution topic>`
Author 1 Paper's Contribution #1 Evidence: `<contribution evidence>`
... (more evidence and contributions)

Based on each of the author's claimed novelties, evidence, and counter-evidence to each other's arguments, you must determine the most meaningful, diverse set of subtopics within the parent topic, "Topic", which best cover the types of contributions each of the papers make. Remember that for each of your selected topics, the papers will be debating which of them makes the better contribution towards the topic. Hence, for each of your subtopics, cite the integer IDs of any relevant contributions from Author 0 or Author 1. At least one of these lists should be non-empty. Overall, our goal is to identify how novel Author 0's paper's contributions towards topic "Topic" are by individually considering their contributions towards your subtopics.

Output your list subtopics (up to k) in the following format: "topic_title": <should be a brief, 10-15 word string where the value is the title of your subtopic>,
"topic_description": <1-2 sentence string explaining the subtopic and what you feel would be most helpful for the papers to debate within the subtopic>,
"author_0_relevant_contributions": <list of integer IDs citing which contribution(s) from Author 0 would be most relevant to this subtopic; can be empty>,
"author_1_relevant_contributions": <list of integer IDs citing which contribution(s) from Author 1 would be most relevant to this subtopic; can be empty>

Table 4: Moderator prompt to generate new topics.

You are a moderator facilitating a debate in which two paper are debating who makes the better contribution towards the following topic:
Topic: `<topic>`
Topic Description: `<topic description>`

`<conversation history between Author 0 and Author 1>`

Below, you are given the previous set of arguments and the current set of arguments.
previous arguments: `<set of arguments before debate>`
current arguments: `<set of arguments after debate>`

You must determine whether progress is being made. DO NOT focus on the language being used. Focus on the content of the arguments.
Specifically, determine the following (True or False for each):
1. progression_of_arguments: Are these arguments sufficiently different enough to necessitate further debate? Are there new, deeper concepts being discussed between the two sets of arguments?
2. meaningful_questions: Within the debate history, each author acknowledges each other's arguments and may ask clarifying questions accordingly. Do you believe that the clarifying questions have not been sufficiently addressed already and would be important to answer through further debate? If there are no questions raised in the debate history by either author, return False.
3. clear_winner: Do you believe that it is clear that one author has won the debate, and it does not need to be further deconstructed (in order to determine which components within each author's contributions are truly better)?

Output your argument in the following format:
"explanation": <2-5 sentence string to explain your reasoning about whether further debate is necessary when comparing the previous arguments and the current arguments>,
"progression_of_arguments": <output a boolean; pick only one of "True" or "False" depending on the history, arguments, and your explanation above>,
"meaningful_questions": <output a boolean; pick only one of "True" or "False" depending on the history, arguments, and your explanation above>,
"clear_winner": <output a boolean; pick only one of "True" or "False" depending on the history, arguments, and your explanation above>

Table 5: Moderator prompt on whether to expand a debate node.

Two authors are debating their respective novelties with respect to the following topic:
Topic: `<Topic>`
Author 0's paper title is: `<Author 0 paper title>`
Author 1's paper title is: `<Author 1 paper title>`
Here is a breakdown of their debates in tree format. At each tree node, we provide the "topic_title" : "topic description", Author 0's corresponding argument and Author 1's corresponding argument:

`<tree (example in Appendix B)>`

Based on the debate breakdown, output a paragraph-long synthesis of the debate which summarizes the similarities and differences between the papers. Structure your summary with initially their similarities (which ideas/aspects overlap between the two papers?) to their differences (what makes the papers unique) in novelties. Focus more on the differences than the similarities.

Table 6: Moderator prompt to summarize the debate into a paragraph.

You are the author of the paper, 'self.paper.title'. The abstract of your work is: <Paper Abstract>.
You are debating another author on the novel contributions your work makes towards the following topic: <Topic>.
Below is a list of relevant evidence retrieved from your paper:<Evidence in the form of sentence excerpts from papers>. Based on the evidence, output a list of 1 to <k> DIVERSE, specific arguments for your position that are all supported by the evidence. Each argument should have a corresponding "argument_title", which is a brief statement of your argument (e.g., Better Efficiency for Training), a "description" explaining your argument and mentioning specific excerpts from your evidence pool, and finally, a list of all "evidence" IDs, which are the integers of the evidence in the input list, that best support your argument. For example, if Evidence #1 and #2 best support your argument, then evidence should be [1,2] (depending on your argument, this list can have more or less than two items). Each argument should make a unique point.

Output your list of arguments in the following format:
"argument_title": <should be a brief, 10-15 word string where the value is the argument_title>,
"description": <1-2 sentence string explaining the argument, including specific excerpts from the evidence pool>,
"evidence": <list of integer IDs citing which evidence from the input list best support your argument>

Table 7: Persona prompt to generate arguments during the debate.

Your objective is to check if a given evidence is relevant to a claim or not (relevancy means evidence that helps either support, refute, or clarify the given claim).
Claim: Argument
Description of Claim: Argument Description
Evidence: Evidence supporting the argument.

Fill out the following schema:
"supports_claim": <"Yes"/"No" if the evidence supports the claim>,
"refutes_claim": <"Yes"/"No" if the evidence refutes the opposition's claim>
"clarifies_claim": <"Yes"/"No" if the evidence clarifies the claim>,
"irrelevant_to_claim": <"Yes"/"No" if the evidence is irrelevant to the claim>,

Table 8: Persona prompt to determine relevant and irrelevant evidences.

You are the author of the paper, '<Paper Title>'. The abstract of your work is: <Paper Abstract>.

You are debating another author (Opposition), whose work is titled, '<Opposition Paper Title>', and abstract is: Opposition Paper Abstract". You are debating the other author on how and why your paper makes a better contribution towards the following topic:
Topic: <topic>
Topic Description: <topic description>

Here are your claimed contributions towards the topic:
Author 0 Paper's Contributions #1: <argument>: <argument description>
Author 0 Paper's Contribution Evidence #1: <evidence towards argument>
Author 1's relevant evidence to potentially counter the quality of this contribution: <counter evidence>
... (more contributions and counter-evidence)

Given the above, make an argument for a specific reason why your contributions towards the topic, Topic: <topic>, are better than the opposition's. If you feel that you do not contribute to the given topic or your contributions ARE NOT better than the opposition's, then state so by conceding to the opposition (e.g., 'I do not believe my paper makes a better contribution than yours') and explain why.

Table 9: Persona prompt to *present* its arguments.

You are the author of the paper, '<Paper Title>'. The abstract of your work is: <Paper Abstract>.

You are debating another author (Opposition), whose work is titled, '<Opposition Paper Title>', and abstract is: Opposition Paper Abstract". You are debating the other author on how and why your paper makes a better contribution towards the following topic:
Topic: <topic>
Topic Description: <topic description>

Here are your claimed contributions towards the topic:
Author 0 Paper's Contributions #1: <argument>: <argument description>
Author 0 Paper's Contribution Evidence #1: <evidence towards argument>
Author 1's relevant evidence to potentially counter the quality of this contribution: <counter evidence>
... (more contributions and counter-evidence)

Here is your conversation debate history with the opposition paper. You must respond to the last argument presented by your opposition in debate. A response may consist of (1) an acknowledgment of the opposition's previous response, (2) answering any of the questions about your paper brought up by the opposition, (3) asking any clarifying questions based on the opposition's claims and reasoning, (4) any clarifications of your own presented arguments based on the opposition, and/or (5) if you feel that the opposition's claim is strong and you do not have sufficient grounds to refute it, then a concession to your opposition.

conversation_history: <conversation history>

Table 10: Persona prompt to *respond* to an argument.

You are the author of the paper, '<Paper Title>'. The abstract of your work is: <Paper Abstract>.

You are debating another author (Opposition), whose work is titled, '<Opposition Paper Title>', and abstract is: Opposition Paper Abstract". You are debating the other author on how and why your paper makes a better contribution towards the following topic:
Topic: <topic>
Topic Description: <topic description>

Here are your claimed contributions towards the topic:
Author 0 Paper's Contributions #1: <argument>: <argument description>
Author 0 Paper's Contribution Evidence #1: <evidence towards argument>
Author 1's relevant evidence to potentially counter the quality of this contribution: <counter evidence>
... (more contributions and counter-evidence)

Based on the debate history and your/your opposition's arguments and evidence, you must construct a new, stronger argument related to the topic. This consists of an argument that addresses/is robust to any doubts or clarifying questions made by the opposition which you feel are valid. If based on the debate, you feel that you do not contribute to the given topic or your contributions ARE NOT better than the opposition's, then state so by conceding to the opposition (e.g., 'I do not believe my paper makes a better contribution than yours') and explain why.

conversation_history: <conversation history>

Table 11: Persona prompt to *revise* to an argument.

B Tree Example

Below, we provide an example of a path that we use in our prompts. Given input topic: Helping Students Fix their Mistakes.

(1) `Level 1 Child Argument`: Personalized Error Resolution Strategies: Debate the effectiveness of adaptive, student-centered approaches in resolving errors, focusing on the role of dynamic question trees and expert decision-making models. This subtopic encourages discussion on the importance of tailoring guidance to individual students' needs and knowledge gaps.

- `Author 0's Argument`: Adaptive Guidance Enhances Personalization via Nuanced Student Modeling. By integrating TreeInstruct's dynamic question tree with a nuanced student modeling framework, our approach provides more effective personalized error resolution strategies than Bridge. This integration enables TreeInstruct to capture the complexities of student thought processes, addressing both independent and dependent mistakes concurrently, while also scaling to accommodate large, diverse student populations. In contrast, Bridge's reliance on predetermined error types and remediation strategies may limit its versatility in complex, multi-bug scenarios.
- `Author 1's Argument`: Expert-Guided Decision-Making for Adaptive, Context-Sensitive Error Resolution. Our approach leverages expert decision-making to inform adaptive, context-sensitive error resolution strategies, demonstrated through the extension of our Bridge model to accommodate complex, multi-bug scenarios. By incorporating nuanced expert thought processes, our method provides more effective personalized guidance than TreeInstruct's adaptive question tree approach. This integration of expert guidance and adaptability addresses the opposition's concerns regarding versatility and scalability in handling diverse student populations.

(a) `Level 2 Child Argument`: Adaptive Guidance in Error Resolution: Debate the effectiveness of adaptive guidance in error resolution strategies, focusing on how each approach tailors feedback to individual students' knowledge states. Discuss the benefits and limitations of each method.

- Author 0's Argument: Adaptive Guidance with Latent Expertise for Personalized Error Resolution. By integrating the strengths of both approaches, our revised argument proposes a hybrid model that combines the adaptive guidance of TreeInstruct with the latent thought processes of expert decision-making. This fusion enables a more nuanced understanding of individual student needs, providing targeted support while maintaining scalability and adaptability. Ultimately, this hybrid approach offers a more effective and personalized error resolution strategy, surpassing the limitations of both adaptive guidance and expert decision-making alone.
- Author 1's Argument: Expert-Infused Adaptive Guidance for Error Resolution. Our revised approach integrates expert decision-making processes into adaptive guidance frameworks, enabling a more nuanced understanding of individual student needs and providing targeted support while maintaining scalability and adaptability. This hybrid model combines the strengths of both approaches, offering a more effective and personalized error resolution strategy. By incorporating expert-informed decision-making into adaptive guidance, we bridge the novice-expert knowledge gap more effectively than either approach alone.