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

Anonymous ACL submission

Abstract

With the exponential growth of research facilitated by modern technology and improved accessibility, scientific discoveries have become increasingly fragmented within and across fields. This makes it challenging to assess the significance, novelty, incremental findings, and equivalent ideas between related works, particularly those from different research communities. Large language models (LLMs) have recently demonstrated strong quantitative and qualitative reasoning abilities, and multi-agent LLM debates have shown promise in handling complex reasoning tasks by exploring diverse perspectives and reasoning paths. Inspired by this, we introduce **Tree-of-Debate** (ToD), a framework which converts scientific papers into LLM personas that debate their respective novelties. To emphasize structured, critical reasoning rather than focusing solely on outcomes, ToD dynamically constructs a debate tree, enabling fine-grained analysis of independent novelty arguments within scholarly articles. Through experiments on scientific literature across various domains, evaluated by expert researchers, we demonstrate that ToD generates informative arguments, effectively contrasts papers, and supports researchers in their literature review.

1 Introduction

Navigating and identifying new and relevant research findings has become non-trivial with the popularity of open-access repositories. For example, arXiv received over 24,000 submissions in October 2024 (arXiv, 2024), inundating researchers with an overwhelming volume of information. This astronomical surge in scholarly articles makes it difficult to identify novel findings and discern the distinctions between related papers, especially those presenting similar ideas from different angles (e.g., papers from different research communities).

Automatically generating comparative summaries of research papers has proven valuable for addressing these challenges (Hayashi et al., 2023). Existing comparative summarization works (Ströhle et al., 2023) typically follow a two-step pipeline: (1) construct extractive summaries for each document to (2) identify their similarities and differences (Lerman and McDonald, 2009; Gunel et al., 2024). However, despite using large language models (LLMs), these methods often focus on surface-level semantic differences, which may not capture the most relevant distinctions. For example, when comparing pre-trained models like "BERT" (Devlin et al., 2019) and "RoBERTa" (Liu et al., 2019), it is crucial to note that RoBERTa omits next-sentence prediction, trains on ten times more data, and achieves superior performance. These insights require complex, comparative reasoning beyond basic semantics, as they rely on understanding BERT's contributions *in the context* of RoBERTa's. Thus, we propose the following principles:

Multi-persona debates elicit complex, comparative reasoning. We explore the use of multi-agent debates for inducing fine-grained, comparative reasoning. These debates simulate group discussions where agents suggest diverse answers, critique one another, and refine responses to produce better outputs (Chan et al., 2023; Liang et al., 2024). Recent work has also introduced defining LLM agents as personas with distinct characteristics or values, enabling them to generate outputs that reflect the diverse perspectives needed to solve multi-faceted problems (Tseng et al., 2024; Wang et al., 2024b). Inspired by this, we propose *converting scientific papers into personas that debate* each other to foster critical analysis. For instance, while the papers debate their respective contributions to a topic, they critically evaluate each other's novelty and significance *relative to their own claims*.

Tree-structured debates allow for independent assessments of different contributions at varying depths. A scientific paper often makes contributions (e.g., methodology, dataset, evalua-

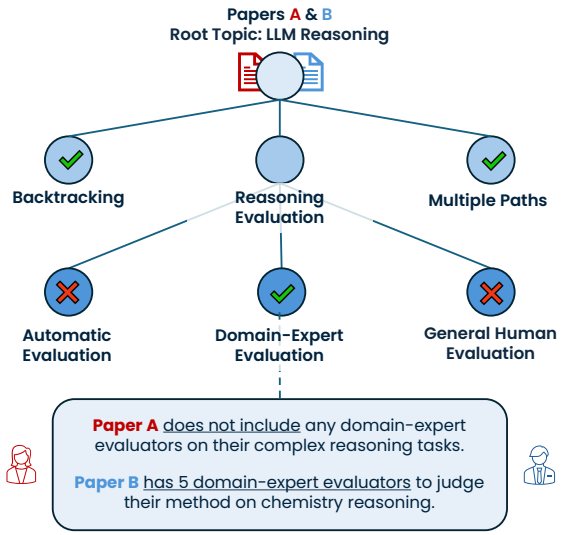

Figure 1: A hierarchy of contributions made by Papers A and B, specific to the root topic. Green check mark: a single paper makes a unique contribution; red X: an overlapping contribution.

tion metric) that can be deconstructed into multiple "sub-ideas." Some sub-ideas may or may not be novel (e.g., uses an existing architecture, but proposes novel fine-tuning and evaluation mechanisms) and consequently, should be independently evaluated for their *degree* of novelty. Hence, an unstructured debate combining all ideas is insufficient for handling the complexity of scientific comparative analysis. We instead propose a *tree-structured* debate, where each *node* represents a specific contribution topic being debated, and an *edge* indicates unresolved points or interesting questions from the parent debate node which warrant further exploration in a child node. Figure 1 illustrates these topical relationships.

Iterative retrieval throughout a debate improves fine-grained reasoning. Due to their lengthy nature, providing an entire paper in-context is ineffective, as details specific to the debate node topic may be overshadowed (Li et al., 2024). Conversely, using only the title and abstract results in high-level comparisons based on surface-level semantic differences. To address these long-context challenges, we propose an *iterative retrieval* process, where retrieval queries are dynamically determined by the debate's content. This ensures the retrieved content is targeted to the specific contribution in question, enabling personas to generate more compelling affirmative or opposing arguments. For instance, as the debate progresses from "reasoning evaluation" to "domain-expert evaluation" in Figure 1, the evidence pool is updated to be more fine-grained and relevant to the subtopic.

We integrate these proposed principles into **Tree-of-Debate**, a framework which dynamically structures a debate between paper personas, conducted by a moderator. First, each persona prepares (**self-deliberation**) by retrieving topic-relevant segments from their paper, identifying their novel contributions, and updating their evidence pool based on the opposition's claimed contributions (Section 3.3). Based on this, the moderator determines the most valuable subtopics to explore (e.g., second level of Figure 1). For each subtopic, a **child debate node is formed**, where each persona presents their initial arguments, responds to one another (e.g., clarifying questions, doubts), and revises their argument based on the interaction. Based on the debate, the moderator determines if the **debate node is worth expanding** and exploring deeper into (Section 3.4). If so, a more fine-grained set of subtopics are determined for the next level of debate children (Section 3.2.2). Our contributions can be summarized as:

- We introduce **Tree-of-Debate**, a structured multi-persona debate 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.

Reproducibility: We provide our dataset and source code[1] to facilitate further studies.

2 Related Work

2.1 Persona Creation & Debate

Similar to how a person's background shapes their abilities, recent work has explored assigning personas to LLMs to capture diverse perspectives and extract unique capabilities (Fu et al., 2023). For instance, Portenoy et al. (2022) creates author personas for author recommendation by applying named entity recognition to papers and matches authors based on shared terminology. As we aim to highlight specific novelties and incremental contributions between two papers, we instead define a *paper persona*. While Portenoy et al. (2022)'s

[1] https://anonymous.4open.science/r/tree-of-debate-5961/README.md

personas represent the author's multiple works and are used solely for comparison, ours represent the paper, actively debating for and defending it. Other studies have also leveraged persona-driven debate by assigning multiple personas– such as affirmative and negative debaters along with a judge— to synthesize diverse reasoning steps for tasks like commonsense generation and arithmetic, thereby reducing confirmation bias inherent in self-reflection methods (Liang et al., 2024). Although our objective differs, we similarly employ debate— not to serve as a means to improve the final output but, rather, as the outcome itself— using the tree directly to generate refined summaries of differences between research papers.

2.2 Comparative Summarization

Generating comparative summaries is challenging due to the diverse ways that differences between two entities can be represented. Traditional graph-based methods (Chen et al., 2022; Ströhle et al., 2023) classify sentences as a claim, similarity, or difference and score them to produce extractive summaries. While we use extractive summarization questions for self-deliberation, ultimately we aim to generate an *abstractive* summary that synthesizes the debate results. More recent works (Luu et al., 2021) fine-tune models to generate explanation sentences by first extracting in-text citation sentences that compare a principal document with a cited one, then maximizing the probability of generating the explanation given the two documents; however, this approach typically yields only a single sentence, which may not fully capture the nuanced differences between papers.

2.3 Generation of Related Works Sections

Multi-document summarization consolidates information from various sources, a task that grows in importance as scientific literature expands (Chen et al., 2022). Certain works within the HCI space (Palani et al., 2023; Lee et al., 2024), which have designed off-the-shelf, interactive systems. However, from a methodological standpoint, one approach (Shi et al., 2023) expands a paper's abstract into semantically similar sentences to form search queries for retrieving relevant papers, and then uses in-context examples to generate related work sections. On the other hand, DIR (Wang et al., 2024a) employs a structured fine-tuning process by prompting a language model to extract commonalities and differences from candidate summaries compared to a gold standard. However, these methods face limitations: the former is highly dependent on the quality of its in-context example without clear guidelines on how to structure the related work. The latter requires fine-tuning the model for each dataset and relies on similarity matching that can restrict summary content (Liu et al., 2023; Hayashi et al., 2023). In contrast, our method uses debate rounds to guide the summary structure, operates at inference-time without training– making it domain-agnostic– and leverages the reasoning capabilities of language models to identify isomorphic properties of ideas beyond mere semantic similarity.

3 Methodology

TREE-OF-DEBATE aims to determine and compare the fine-grained scientific claims of two papers through a methodology inspired by formal debate. Our overall framework is presented in Figure 2.

3.1 Preliminaries

3.1.1 Problem Formulation

We assume two papers, p_1 and p_2, and a topic n_0 (e.g., "inference-time LLM reasoning methods") are provided as input by the user. Our goal is to determine the specific novelties, incremental additions, and equivalent contributions relevant to n_0 between p_1 and p_2, producing a **debate tree** T with a corresponding **comparative summary** S. In T, each node n_i represents a topic (with n_0 as the root), where topic n_i guides the specific debate occurring at that node. Topics may pertain to both papers or only one (e.g., in Figure 1, only $p_i = B$ includes "Backtracking"). An edge from n_i to n_j^i indicates that n_j is a subtopic of n_i that merits further exploration.

3.1.2 Segment-level Retrieval

An effective debate is contingent on an individual's preparation *before* the debate and their ability to retrieve knowledge dynamically *during* the debate. We employ a retrieval embedding model (Xiao et al., 2023) and cosine-similarity to compute and rank segment-level embeddings. We chunk each paper into roughly three-sentence segments such that it is easily comprehensible during the debate.

3.2 TREE-OF-DEBATE Setup

We conduct a multi-persona debate between two paper personas, p_1 and p_2, based on the high-level claim, *p_1 is better than the p_2 for topic n_i*. Our

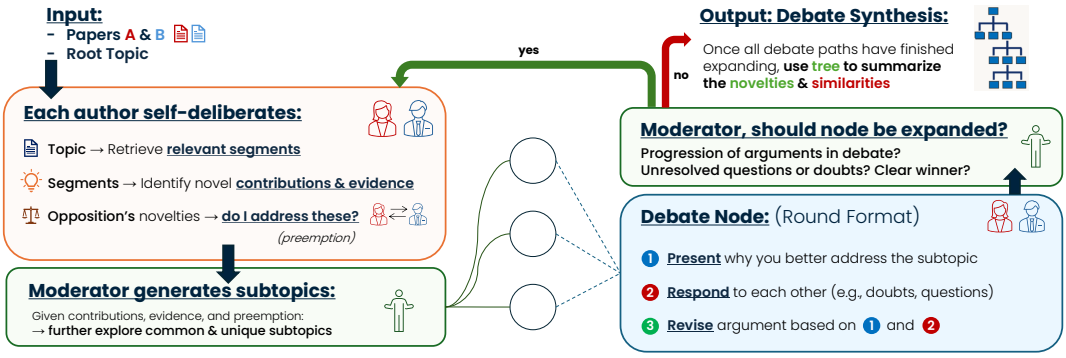

Figure 2: We propose Tree-of-Debate, a novel framework which guides a multi-persona debate using a retrieval-augmented tree. $A \rightarrow B$ within the diagram translates to the statement, "Given A, a persona arrives at B".

goal is not to determine a final debate "winner" but to capture the **specific reasoning *induced*** by the debate format, reflected in the progression of arguments and the degree of novelty in each paper's claims. While we explored novelty-specific claims (e.g., "p_1's contribution towards n_i is more novel than p_2's"), this led to more surface-level arguments. Moreover, a paper often features a breadth of claims/ideas, motivating a debate structure that is flexible to explore these different angles independently. Thus, we propose a **TREE-OF-DEBATE** (ToD), T, where each node represents a round of debate (Section 3.2.2). A directed edge from parent node n_p to child node n_c^p indicates the debate progressing into one of n_p's subtopics n_c^p (out of potentially k subtopics). Leaf nodes indicate no further progression in argumentation is evident.

3.2.1 Constructing the Personas

We leverage an LLM agent to embody each debate persona, allowing for retrieved information from the papers and the debate history to be easily integrated into its context:

- **Papers:** Each paper persona is given the title, abstract, and retrieved segments relevant to the starting topic n_0 (updated at each self-deliberation stage (Section 3.3)). Each paper persona's p_i role is to argue that their contributions towards the topic n_i are better than persona p_j's.

- **Moderator:** Using the same underlying model as the paper personas, the moderator (i) identifies key debate subtopics for determining the papers' similarities and differences, (ii) judges the progression of the debate based on the authors' arguments, and (iii) synthesizes the debate tree into a comparative summary.

3.2.2 Tree Node Format

Each tree node with topic n_i undergoes a three-stage debate (pre-, during, and post-debate). The format is as follows:

1. **Self-Deliberation (*Section 3.3*)**: Each persona p_a retrieves segments S_i^a relevant to n_i, generates k claims C_i^a for their novel contributions, cites corresponding evidence $E_i^a \subseteq S_i^a$, and collects counter-evidence from their own paper $\widetilde{E}_i^a$. The moderator then selects k new subtopics for the next level of k children, $n_j^i \in N^i$.

2. **Debate (Prompts 13, 14, and 15)**: For each child debate node n_j^i, each persona p_a (i) presents an argument that p_a is <u>better</u> than p_b on n_j, (ii) responds to the opposing argument, and (iii) revises their argument accordingly.

3. **Determine Expansion *(Section 3.4)***: Based on the debate at n_j^i, the moderator determines whether the arguments progressed or introduced any unresolved questions meriting another round. If so, the moderator triggers self-deliberation for $n_i \rightarrow n_j^i$.

3.3 Self-Deliberation

Self-deliberation is an argumentative strategy (Tindale, 2020) that enables one to "argue with oneself" by considering alternative views, aiming to arrive at the best, most well-justified conclusion. We integrate self-deliberation into our multi-persona debate for a given topic node n_i and paper $p_{a \in \{1,2\}}$:

1. Retrieve relevant segments S_a^i from p_a that are closely related to n_i.

2. Generate k claims $c_j \in C_i^a$ on the *novel contributions* of p_a toward n_i. Each claim includes a title, description, and a set of mapped evidence $E_{(i,j)}^a \subseteq S_i^a$ (see Prompt 11).

3. Preempt p_b's contributions, where C_i^b is exposed in-context to p_a and p_a retrieves another round of evidence $\widetilde{E}_i^a$ aimed at targeting p_b's claims.

4. The moderator then uses E_i, $\widetilde{E}_i$, and C_i to generate a list of **subtopics** for further exploration.

Retrieving relevant segments. For each paper $p_{a\in\{1,2\}}$, we retrieve the top δ segments $S^i_{a\in\{1,2\}}$ conditioned on node topic n_i using the retrieval embedding model (Xiao et al., 2023) (Section 3.1.2). We embed n_i using the query format: "`[topic name] : [topic description]`" (the moderator generates a description for each non-root node). These segments form two separate pools of evidence per paper to compose their novelty claims.

Preemption. For each of paper p_b's novelty claims $c^b_{(i,j)} \in C^b_i$ with its corresponding evidence E^b_i, we retrieve additional segments $\widetilde{E}^a_{(i,j)}$ from p_a using the concatenated title and description of $c^b_{(i,j)}$ as the query. Each retrieved segment e is then filtered using an LLM-based step (Prompt 12) that evaluates whether e (1) *supports*, (2) *refutes*, (3) *clarifies* p_b's claim, or (4) is *irrelevant*. While redundant, we notice that explicitly including (4) as an option helps with filtration performance. If either (1-3) are true or (4) is false, then e is filtered out. If $|\widetilde{E}^i_{(a,j)}| = 0$, we indicate that p_a does not address p_b's claim, $c^b_{(i,j)}$. Overall, preemption allows the paper personas to be better prepared for their opposition's arguments ahead of the debate.

Subtopic Generation. Using each paper's title, abstract, C_i, E_i, and $\widetilde{E}_i$, the moderator generates k **subtopics** $n^i_j \in N^i$ that should be further explored (Prompt 8). The moderator maps each subtopic to at least one claim c_i from either p_1 and/or p_2, forming child debate nodes that explore overlapping topics (e.g., "Reasoning Evaluation" in Figure 1) or topics potentially unique to one paper (e.g., "Multiple Paths").

3.4 Debate Tree Expansion & Synthesis

While we motivate the personas to examine and debate whether their work proposes a *better* idea than their opposition, this mechanism is intended to (1) emphasize the ***reasoning behind the idea*** and (2) provoke debate on the ***novelty behind the ideas***, relative to each other. In other words, we hypothesize that two very similar ideas (e.g., "Reasoning Evaluation" in Figure 1) will typically lead to a *longer* debate subtree on which approach is better. Conversely, a uniquely novel approach or task proposed in p_1, relative to p_2, may result in a *shorter* debate as the moderator will ideally determine that p_2 does not address p_1's claim in their work and thus has a weak argument (e.g., "Backtracking" in Figure 1). To facilitate this process, the moderator's core tasks are detailed below.

3.4.1 Determining Round Depth Expansion

For debate node n_i, the moderator assesses the following (Prompt 9):

1. **Argument Progression**: Is there sufficient evolution in the arguments or new, deeper concepts being introduced to justify further debate?

2. **Meaningful Questions**: Have clarifying questions been raised that remain unanswered and merit further discussion? If no questions are raised, the moderator returns False.

3. **Clear Winner**: Is it clear that one paper has won the debate, as their contributions are truly better and do not warrant deconstruction (to determine which *subcomponents* are truly better)?

If either (1) or (2) holds true, or if (3) does not indicate a clear winner, the moderator proceeds to the self-deliberation stage (Section 3.3) to identify new subtopics and expand n_i to $n^i_j \in N^i$. Otherwise, expansion stops for that debate path. A maximum tree depth is also enforced.

3.4.2 Debate Synthesis

Once ToD has converged (i.e., all debate paths have been adequately expanded), the moderator synthesizes the entire debate tree into a paragraph-long comparative summary. The debate tree is provided in-context, with each node $n_i \in T$ containing the following information: node topic **title**, node topic **description**, persona p_1's **revised argument** (at the end of the debate round), and persona p_2's **revised argument**. The synthesis should first explain the papers' novelty similarities and then detail their differences, with greater emphasis on the latter. We provide the prompt in Prompt 10 and an example subtree in Appendix H.

4 Experimental Design

We choose `Llama-3.1-Nemotron-70B-Instruct-HF`, an *open-source model*, as the base model for all experiments. We sample from the top 1% of the tokens and use the same temperature settings across all samples (details on setting and hardware provided in Appendix A).

4.1 Dataset

No datasets currently exist for comparing *non-citing* pairs of scientific papers— an overlooked setting, especially given the growing scale of literature where not all relevant work can be cited. Consequently, we aimed to construct a dataset with papers that both cite and do not cite each other, in order to test the robustness of Tree-of-Debate. However, novelty comparison between papers is a highly specialized and expensive task, requiring rich domain expertise to verify– especially if such papers do not explicitly cite one another. Thus, we gathered five domain expert researchers (detailed provided in Appendix D) to construct a dataset of 100 paper pairs across natural language processing, data mining, electrical engineering, and aerospace engineering (further details provided in Appendix E). Each researcher identified at least five papers they were highly familiar with, such that they could perform a detailed and informed human evaluation. They were instructed to annotate each pair with a root topic and whether: (1) the papers roughly focus on the same task but differ in **methodology**, or (2) they work on different **tasks** that are applied to *similar* motivations. Furthermore, they noted if the papers explicitly **cited** each other **or not**. Table 1 shows the dataset distribution.

Category	Method	Task	Total
Cited	15	15	30
Not Cited	30	40	70
Total	45	55	100

Table 1: # of paper pairs/summaries per category.

4.2 Baselines

Given that our primary goal is to demonstrate the difference in inference-time comparative reasoning capabilities between ToD and current LLMs, we design the following prompting-based baselines: (1) **Single Stage** uses the title, abstract and introduction sections of both papers to directly generate a comparative summary (Martin-Boyle et al., 2024); (2) **Two Stage** first individually summarizes each paper based on the title, abstract and introductions, and then uses the generated summaries to generate a comparative summary (Zhang et al., 2024).

To contextualize improvements from each component in ToD, we construct the following ablative methods: (1) **ToD (No Tree)** removes the tree structure by merging child arguments into one and considering the combined subtopic as the debate node topic; (2) **ToD (No SD)** removes self-deliberation (SD) to test the impact of iterative retrieval based on debate progression. No SD relies on the title, abstract, and introduction of each paper instead of retrieving based on the subtopic. We use the same LLM for all comparisons. Complete baseline and ablation details are provided in Appendix B.

4.3 Evaluation Metrics

The same domain-experts from Section 4.1 manually evaluate each of their chosen pairs in-depth, assessing various qualities of the 100 summaries. We normalize each of the scores below and scale them by 100 for the final results in Table 2 (full metric guidelines provided in Appendix C):

- **Factuality**: *How factual is the summary?* Each sentence is given a 1/0 binary score for factuality, and the scores are averaged across the summary.
- **Breadth**: *Is the summary comprehensive and complete?* Each summary is rated from 0-4 ("*not at all*" to "*very*").
- **Contextualization**: *Does the summary explain and/or justify the posed similarities/differences between the papers, as opposed to just mentioning them?* Each summary is rated from 0-4 ("*not at all*" to "*very*").

5 Experimental Results

Overall Performance & Analysis. Table 2 shows the performance of Tree-of-Debate (ToD) compared with the baselines on factuality, breadth of comparison (completeness), and contextualization. We observe that the domain-experts found ToD summaries 6.85% more complete and 25.98% more contextualized compared to the most competitive baseline. This observation indicates that multi-persona debate trees help analyze pairs of papers to uncover more fine-grained contributions, as well as identifying *connections* between the papers. Given that all samples were carefully annotated and evaluated by domain-experts, we are able to draw several interesting insights, which we list below:

Structured debate is necessary for eliciting contextualized comparative summaries. Our results show that TREE-OF-DEBATE significantly improves contextualization, achieving an average score of 95.21% across all settings, compared to 75.57% for the strongest baseline (Two-Stage). Our domain-expert evaluators frequently observed that the LLM-generated summaries (Single and Two Stage) at face value mention a breadth of

Table 2: We showcase TREE-OF-DEBATE's strong performance across all comparison settings. **Bolded** values indicate the top score; underlined indicates second-highest. We include two ablations which remove the tree structure (*No Tree*) and self-deliberation (*No SD*), respectively.

Baseline	Method + Cite			Method + Not Cited			Overall (Method)		
	Breadth	*Context*	*Factuality*	*Breadth*	*Context*	*Factuality*	*Breadth*	*Context*	*Factuality*
Single Stage	93.33	80.00	**93.48**	89.81	73.15	94.90	91.07	75.59	94.39
Two Stage	85.00	71.67	91.42	87.03	72.22	94.74	86.31	72.02	93.69
Tree-of-Debate	**96.67**	**93.33**	92.80	**93.52**	93.52	**96.80**	**94.64**	**93.45**	**95.37**
ToD (No Tree)	81.67	63.33	91.25	79.63	70.37	89.46	80.36	67.86	89.82
ToD (No SD)	91.67	76.67	82.33	86.11	**94.44**	76.33	88.10	88.10	78.21

Baseline	Task + Cite			Task + Not Cited			Overall (Task)		
	Breadth	*Context*	*Factuality*	*Breadth*	*Context*	*Factuality*	*Breadth*	*Context*	*Factuality*
Single Stage	81.82	65.91	94.44	88.46	71.15	92.31	87.00	70.00	92.78
Two Stage	89.29	82.14	**97.14**	86.84	78.29	**94.18**	87.22	78.89	**94.54**
Tree-of-Debate	92.86	**100**	87.89	**96.05**	**96.05**	94.02	**95.19**	**97.11**	92.36
ToD (No Tree)	67.86	67.86	61.11	83.55	73.03	91.43	81.11	72.22	89.74
ToD (No SD)	**96.43**	96.43	75.00	81.58	94.08	82.27	83.89	94.44	81.79

Table 3: Debate Round for Tree of Thoughts (ToT) vs. Chain-of-Thought (CoT); turns are summarized.

Stage	Dialogue for Subtopic: "Generalizability and Flexibility of Prompting Methods"
Present	**ToT**: ToT *enhances flexibility* via deliberate reasoning by exploring multiple paths, self-evaluation of choices, and lookahead/backtracking capabilities– shown on a diverse set of tasks like Game of 24 and Creative Writing. This is in contrast to CoT's lack of deliberate reasoning and adaptability.
	CoT: CoT excels in *generalizability through its simplicity*, requiring minimal examples to achieve strong results across tasks. Unlike ToT, it *avoids complex setups and can generalize to unseen tasks*, making it more adaptable with fewer computational demands.
Respond	**ToT**: While CoT is simpler, ToT's added complexity is a necessary *trade-off for deeper reasoning tasks*, requiring non-trivial planning and search. CoT's ability to handle deeply nested or conditional reasoning remains unproven.
	CoT: ToT may help in specific tasks, but CoT's simplicity deliberately ensures *broader applicability* without overfitting to specific task requirements. Though direct evidence for nested reasoning is lacking, CoT's state-of-the-art GSM8K performance suggests complex reasoning robustness. A future study can compare both methods on nested reasoning to settle this question empirically.
Revise	**ToT**: ToT has *enhanced flexibility via adaptive reasoning path exploration*– exploring multiple reasoning paths and leveraging self-evaluation and lookahead/backtracking to tackle complex tasks– as shown on our diverse problem domains. Our deliberate decision-making is advantageous for tasks with *deeply nested or highly conditional reasoning paths*, where CoT may falter.
	CoT: CoT *maintains strong generalizability by balancing simplicity and effectiveness*– including those with complex reasoning requirements– *without overfitting* to specific task needs. This is demonstrated through its SoTA performance on GSM8K.

specific contributions made by each paper, noting them as either similarities or differences. However, these tend to resemble ***extractive*** summaries, where phrases that are *semantically similar or dissimilar* are extracted from the papers' abstracts and introduction and are posed as similarities and differences respectively, with no context provided on *why* this is the case. We further analyze this finding in Section 5.1. Moreover, we see that without the tree-structure, the debate's analysis quality deteriorates. Specifically, for our "No Tree" ablation study (Table 2), we modify our method to combine all proposed subtopics by the moderator (Section 3.3) into a single high-level topic. This leads to lower contextualization and breadth of contributions discussed, due to reasoning difficulty in disentangling the contributions during the debate. Thus, *structuring the debate is necessary to experience its contextualization benefits*.

Retrieval-augmented debate ensures factuality and breadth of comparison. In Table 2, we compare the performance between ToD and "No SD", where the latter ablation replaces the iterative retrieval step with simply providing the paper's title, abstract, and introduction as in-context evidence across the entire debate tree (same setting as the two baselines). We observe that No SD still experiences the strong benefits of the structured de-

Model	Comparative Summary Excerpts
Two Stage	MEGClass focuses on mutually-enhancing text granularities, iteratively refining its model through feedback between document, sentence, and word levels, and *requires an initial set of weakly labeled data*. In contrast, LOTClass *innovates by relying solely on label names* to train its model, employing a language model self-training approach that *obviates the need for any labeled documents*, achieving high accuracy without human-labeled data.
ToD	*Both papers leverage limited labeled data* to achieve robust performance. Both MEGClass and LOTClass also employ *iterative refinement techniques*, with MEGClass using iterative feedback and LOTClass utilizing self-training on unlabeled data. MEGClass *uniquely leverages mutually-enhancing text granularities*, capturing both coarse- and fine-grained context signals to provide a more comprehensive understanding of class labels. In contrast, LOTClass relies *solely on label names*, leveraging pre-trained language models to drive a self-training mechanism. Furthermore, MEGClass's adaptive granularity balancing approach provides robustness to label name selection, whereas LOTClass's *reliance on label names alone may introduce biases.*

Table 4: Case study on two weakly supervised text classification works (Kargupta et al., 2023; Meng et al., 2020).

bate, evident through its similarly high contextualization score (91.37%, compared to ToD's 95.47% and Two Stage's 75.57%). We note a drop in the factuality and breadth of the summary. The information present in the abstract and introduction is not detailed or deep enough to facilitate fine-grained discussion of the paper. This leads to more noise, and hallucinations, which negatively impacts No SD's breadth and factuality scores. Therefore, *iterative evidence retrieval is necessary for exploring in breadth and minimizing hallucinations*.

ToD is robust to all comparison settings. We analyze four different comparison settings (as detailed in Section 4.1), where two papers may or may not cite each other and differ in either their *task* or *method*. We can see through Table 2 that regardless of their comparison type, ToD demonstrates consistently high performance– with only an average standard deviation $\sigma = 2.49$ across breadth and soundness, while Single Stage's $\sigma = 5.32$ and Two Stage's $\sigma = 3.39$. While there are no consistent trends across the different settings, studying them allows us to *ensure the consistency of ToD*.

5.1 Qualitative Case Study

Evolution via Critical Reasoning. Our approach enables paper personas to refine their comparisons by addressing counterarguments elicited through debate. Table 3 illustrates this through a debate round between Tree of Thoughts (ToT) and Chain-of-Thought (CoT) on their generalizability and flexibility. ToT initially highlights its *flexibility* through deliberate reasoning, exploring multiple paths with lookahead/backtracking. However, CoT counters that ToT's complexity hinders broad applicability, whereas CoT can easily generalize to a multitude of tasks with minimal setup while still achieving state-of-the-art results. In the revision stage, both refine their arguments: ToT emphasizes adaptive exploration for *complex reasoning*, while CoT maintains its simplicity, now citing empirical success in complex tasks. It is interesting to note that during the *Respond* stage, CoT ***suggests a future study*** that would help validate the claims made in the debate.

Contextualized Summaries. The baselines' summaries tend to be extractive, not explaining why certain comparisons are important and also mistaking similarities for differences if their wording is semantically dissimilar. Table 4 demonstrates this finding; Two Stage *mistakenly states* that MEGClass "requires an initial set of weakly labeled data" while LOTClass "innovates by relying solely on label names". However, this is false– both methods only require the class label names. This similarity is identified by ToD, which uses it to contextualize the methods' differences: both use label names, however MEGClass considers all three levels of text granularity, while LOTClass solely relies on the label names. ToD also provides *further insight* into LOTClass's over-reliance on label names potentially introducing biases. ToD's debate format elicits critical reasoning and allows for a more insightful, contextualized comparison. Appendices H and F include in-depth qualitative analyses on an additional sample's tree and output summaries.

6 Conclusion

Automatic summarization is essential for managing the growing volume of research. We introduce **Tree-of-Debate**, a structured approach that models papers as personas engaging in a debate to extract their key similarities and differences. Our method organizes debates into a hierarchical structure to produce abstractive and contextualization summaries while preserving their factuality. With thorough domain-expert empirical evaluation and qualitative case studies, we demonstrate that Tree-of-Debate significantly outperforms baselines.

7 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.

We also observe a ***slight trade-off between contextualization and factuality***. Especially in more challenging comparisons where the overlap between the papers may not be explicit (e.g., task-based comparisons), ToD must take more of a leap with respect to its reasoning to identify similarities and differences between the papers, which may lead to further hallucinations. Furthermore, we notice that as a debate path progresses to deeper levels, if certain evidence is not present to support a paper's fine-grained claims, the personas begin to suggest potential future studies or even new methods (e.g., combining certain strengths of the two methods). While this does introduce some noise (as we can see through our competitive factuality scores, this is minimal) to our output summary, these types of "hallucinations" present exciting new paths for research to explore.

Tree-of-Debate can be ***extended 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.

References

arXiv. 2024. arxiv monthly submission statistics. https://arxiv.org/stats/monthly_submissions. Accessed: 2024-10-13.

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*.

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*, pages 5978–5988, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

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*.

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*, pages 1019–1031, Dubrovnik, Croatia. Association for Computational Linguistics.

Priyanka Kargupta, Tanay Komarlu, Susik Yoon, Xuan Wang, and Jiawei Han. 2023. MEGClass: Extremely weakly supervised text classification via mutually-enhancing text granularities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10543–10558, Singapore. Association for Computational Linguistics.

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*.

Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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*, pages 113–116.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

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*, pages 5351–5355.

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.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. *Preprint*, arXiv:2002.00317.

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.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Jason Portenoy, Marissa Radensky, Jevin West, Eric Horvitz, Daniel Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. *Preprint*, arXiv:2108.05669.

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*, pages 5785–5799, Singapore. Association for Computational Linguistics.

Thomas Ströhle, Ricardo Campos, and Adam Jatowt. 2023. Contrastive text summarization: a survey. *International Journal of Data Science and Analytics*, 18(4):353–367.

Christopher W Tindale. 2020. Self-deliberation and the strategy of the pseudo-dialogue. *Co-herencia*, 17(32):159–178.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Pancheng Wang, Shasha Li, Dong Li, Kehan Long, Jintao Tang, and Ting Wang. 2024a. 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*, pages 2028–2037.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024b. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

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.*, 61:103696.

A Experimental Settings

We choose `Llama-3.1-Nemotron-70B-Instruct-HF`, an *open-source model*, as the base model for all experiments. 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 based on the nature of the given task (same setting across all samples):

- Paper generates arguments: 0.3
- Paper whether evidence is relevant: 0
- Paper presents its argument: 0.1
- Paper responds to the opposition's argument: 0.4
- Paper revises its argument: 0.4
- Moderator generating subtopics: 0.3
- Moderator determines whether to expand the debate note: 0.1
- Moderator summarizes a debate path: 0.4

We set the number of retrieved segments $\delta = 5$ so that we can gather a sufficient amount of evidence while not overwhelming the debate with long-context. We set the number of generated subtopics $k = 3$, for covering a reasonable breadth of topics while minimizing redundancy. Finally, we set the maximum debate tree depth $l = 3$ for adequate exploration. We use vLLM (Kwon et al., 2023) for distributed and constrained generation on four `NVIDIA A100` GPUs.

B 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 (Martin-Boyle et al., 2024).
- **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 (Zhang et al., 2024).

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.

C Evaluation Metrics

The same domain-experts from Section 4.1 manually evaluate each sample in-depth, assessing various qualities of the summaries:

- **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.
- **Completeness**: *Is the summary comprehensive and complete?* This is evaluated using the following Likert scale:
  0. Not at all, the summary misses (MULTIPLE) major points.
  1. No, the summary misses a (SINGULAR) major point.
  2. Somewhat, the summary misses minor points.
  3. Yes, the summary covers the major points, but still is not what I would expect.
  4. Very comprehensive, the summary covers the major points.
- **Contextualization**: *Does the summary explain and/or justify the posed similarities/differences between the papers, as opposed to just mentioning them?*
  0. Not at all, the summary is simply extractive– it just seems to take different subtopics from each paper and doesn't synthesize them– no justification behind similarities and differences.
  1. No, the summary attempts at some level of synthesis, but it is not meaningful.

2. Somewhat, the summary attempts at synthesizing at most one point.
3. Yes, the summary contains meaningful synthesis but only for a minority of points.
4. Strongly contextualized, the summary contains meaningful synthesis across all major points.

D Domain-Expert Profiles

Given that novelty comparison between papers is a highly specialized and expensive task, we gather five domain experts to both collect and annotate our dataset, as well as evaluate Tree-of-Debate's generated summaries based on their respective samples. Each domain expert is a graduate student with 3+ years of research experience in a specialized area:

1. **Domain Expert #1:** A third-year PhD in computer science with ten publications; research expertise is text mining and data mining.
2. **Domain Expert #2:** A third-year PhD in aerospace engineering with two publications; research expertise is in electric propulsion.
3. **Domain Expert #3:** A second-year PhD (with two years of a research-track Masters) in electrical engineering with four publications; research expertise is in in-memory computing and wireless communications.
4. **Domain Expert #4:** A first-year PhD (with two years of a research-track Masters) in computer science with six publications; research expertise is data-efficient natural language processing.
5. **Domain Expert #5:** A first-year PhD (with two years of a research-track Masters) in computer science with twenty-five publications; research expertise is large language model training and efficiency.

E Dataset Specifications

As mentioned earlier, TREE-OF-DEBATE's dataset contains 100 samples, and Table 1 specifies the breakdown. Each sample contains the following:

(a) Topic: a short, vague description of the theme of the two papers
(b) Paper #1 arXiv Link
(c) Paper #1 Title
(d) Paper #1 Abstract
(e) Paper #1 Introduction
(f) Paper #2 arXiv Link
(g) Paper #2 Title
(h) Paper #2 Abstract
(i) Paper #2 Introduction
(j) Method/Task: 0 if the papers differ in methodology (but have the same task) and 1 if the papers differ in the task (but the methodology is generally the same)
(k) Cite/No: 0 if the papers do not cite each other, and 1 if the papers cite each other.

We provide the dataset (as a tab-separated file) in our code repository: https://anonymous.4open.science/r/tree-of-debate-5961/dataset/tree_of_debate_dataset.tsv. 5 contains a few rows from the dataset.

F Qualitative Case Study: TreeInstruct vs. BRIDGE

Tables 6 and 7 contain 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,

Table 5: Snapshot of 5 rows of the dataset. The table is transposed in order to display it on the full page. Notice the diversity in topics, not just in machine learning, but also in other disciplines.

Index	1	2	3	4	5
Topic	helping students understand their mistakes and misunderstandings	Continual pretraining of Bert for retrieval tasks	contrastive learning on graphs	hall thruster erosion	massive MIMO baseband processing
Paper #1 arXiv	https://arxiv.org/pdf/2406.11709	https://arxiv.org/pdf/2004.12832	https://arxiv.org/pdf/2103.00113	https://deepblue.lib.umich.edu/bitstream/handle/2027.42...	https://arxiv.org/pdf/2407.06755
Paper #1 Title	Instruct, Not Assist: LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging	ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT	Anomaly Detection on Attributed Networks via Contrastive Self-Supervised Learning	An Investigation of Factors Involved in Hall Thruster Wall Erosion Modeling	A 46 Gbps 12 pJ/b Sparsity-Adaptive Beamspace Equalizer for mmWave Massive MIMO in 22FDX
Paper #1 Abstract	Socratic questioning is an effective teaching...	Recent process in Natural Language Understanding...	Anomaly detection on attributed networks attracts	A hydrodynamic description of the plasma flow...	We present a GlobalFoundries 22FDX FD-SOI...
Paper #1 Introduction	With the rapidly expanding conversational...	Over the past few years, the Information...	Attributed networks (a.k.a. attributed graphs...	As lifetime requirements desired for Hall thruster...	Fifth generation (5G) and beyond-5G wireless...
Paper #2 arXiv Link	https://arxiv.org/pdf/2310.10648	https://arxiv.org/pdf/1810.04805	https://arxiv.org/pdf/2310.14525	https://electricrocket.org/IEPC/IEPC-2007-151.pdf	https://arxiv.org/pdf/1910.00756
Paper #2 Title	Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Graph Ranking Contrastive Learning: A Extremely Simple yet Efficient Method	Investigation of the SPT operation and discharge chamber wall erosion rate under increased discharge voltages	Beamspace Channel Estimation for Massive MIMO mmWave Systems: Algorithm and VLSI Design
Paper #2 Abstract	Scaling high-quality tutoring remain a major challenge...	We introduce a new language representation...	Graph contrastive Learning (GCL) has emerged as...	New results of the thruster operation specifics	Millimeter-wave (mmWave) communication
Paper #2 Introduction	Human tutoring plays a critical role in accelerating student...	Language model pre-training has been shown to be effective...	Graph Neural Networks (GNNs) have become the standard	Nowadays it is reasonable to develop SPT with increased	Millimeter-wave (mmWave) communication [2], [3]
Method /Task	1	1	1	0	1
Cite/No	1	1	0	0	0

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 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

Table 6: Summaries from two methods: ToD (No Tree) – the TREE-OF-DEBATE ablation – 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. The rest of the table follows in 7.

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.

Table 7: Summaries from two methods: ToD (No SD) – the TREE-OF-DEBATE ablation – and TREE-OF-DEBATE. 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. The rest of the table follows in 6.

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.

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.

G 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 8: 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 deconstructured (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 9: 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 H)>

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 10: 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 11: 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 12: 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 13: 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 14: 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 15: Persona prompt to *revise* to an argument.

H 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.