Unlocking Varied Perspectives: A Persona-Based Multi-Agent Framework with Debate-Driven Text Planning for Argument Generation

Anonymous ACL submission

Abstract

Writing persuasive arguments is a challenging task for both humans and machines. It entails incorporating high-level beliefs from various perspectives on the topic, along with deliberate reasoning and planning to construct a coherent narrative. Current language models often generate surface tokens autoregressively, lacking explicit integration of these underlying controls, resulting in limited output diversity and coherence. In this work, we propose a persona-based multi-agent framework for argument writing. Inspired by the human debate, we first assign each agent a persona representing its high-level beliefs from a unique perspective, and then design an agent interaction process so that the agents can collaboratively debate and discuss the idea to form an overall plan for argument writing. Such debate process enables fluid and nonlinear development of ideas. We evaluate our framework on argumentative essay writing. The results show that our framework can generate more diverse and persuasive arguments through both automatic and human evaluations.

1 Introduction

One of the most common formats of opinion-based communication is argumentation, where users present their viewpoints and attempt to persuade others to adopt their stance on various topics. Writing argumentative essays on controversial topics presents significant challenges in natural language processing (Hua and Wang, 2018; Wang et al., 2023a; Hua et al., 2019). The complexity of this task stems from several requirements: Firstly, it necessitates social understanding capabilities for a profound comprehension of the topic and the inclusion of varied, pertinent viewpoints to bolster the argument's persuasiveness. Secondly, it demands strong logical reasoning and strategic text planning to create a coherent overarching structure, which integrates different viewpoints into a well-organized discourse. Lastly, fundamental writing skills are crucial for effectively transforming the plans into surface text.

Recent large language models (LLMs) have demonstrated impressive outcomes (Touvron et al., 2023a; Ouyang et al., 2022; Touvron et al., 2023b; Achiam et al., 2023). Yet, they still face challenges when tasked with writing argumentative essays (Hu et al., 2023). Despite their effectiveness, LLMs often fail to offer diverse and rich content, particularly in generating subjective content with multiple viewpoints (Muscato et al., 2024; Hayati et al., 2023). This limitation arises because LLMs are trained to model averages and may overlook the nuance and in-group variation of perspectives (Sorensen et al., 2024). Nevertheless, such diversity is essential for crafting persuasive arguments that resonate with a wide audience.

Additionally, current LLMs often generate text autoregressively without explicit planning (Bubeck et al., 2023; Wang et al., 2022), contrasting with human writing that typically involves extensive planning to establish a coherent high-level logic flow (Hu et al., 2022; Flower and Hayes, 1981). Recent efforts address this by decomposing the end-to-end generation into content planning and surface writing (Yang et al., 2022; Zhou et al., 2023). While they are effective for narrative texts like stories, planning for arguments is inherently more complex. It requires nonlinear thinking to ensure a solid logical structure, effectively connect diverse perspectives, and proactively counter potential objections.

In this paper, we propose a persona-based multi-agent framework built upon LLMs for writing argumentative essays that are perspective-diverse and logically coherent. Recent work shows that assigning personas to LLMs can enhance the performance towards specific perspective and believable human behavior (Jiang et al., 2023; Xu et al., 2023). To enhance perspective diversity, our framework employs multiple agents, each endowed with a distinct

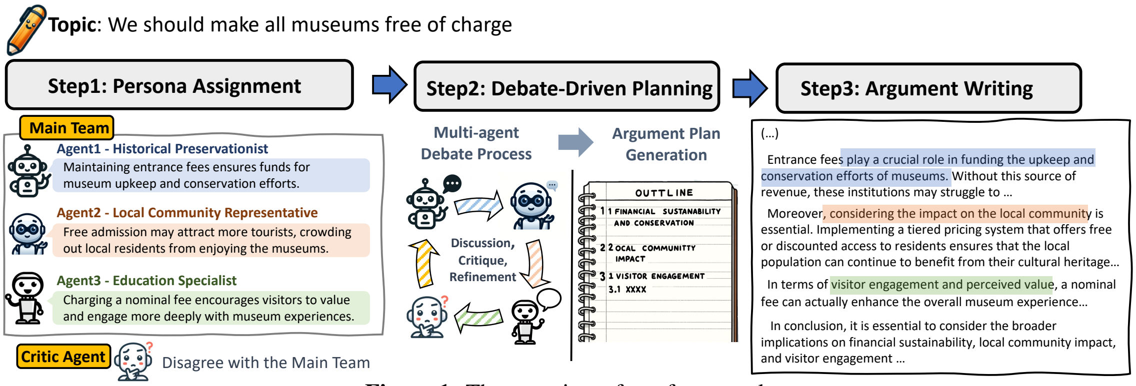

Figure 1: The overview of our framework.

persona, representing a unique viewpoint relevant to the input topic. This multi-persona collaboration brings unique perspectives and expertise to the table, thus crafting a more compelling and persuasive argument (Johnson and Blair, 2006).

Inspired by previous work utilizing multi-agent debate to improve LLMs' performance (Wang et al., 2023b; Du et al., 2023), we model text planning as a debate process among the agents. Additionally, a critic agent is integrated to challenge the idea presented, ensuring a robust discussion. During the debate, agents engage in dialogue, respond to critiques, and progressively refine their ideas. This collaboration not only fosters creativity and critical thinking but also aids in self-revision and self-critic. The discussions are then distilled into an argument plan that offers diverse viewpoints and maintains logical coherence. Unlike previous planning methods that *sequentially* outline content (Hua and Wang, 2020; Goldfarb-Tarrant et al., 2020; Yang et al., 2022), our debate-driven planning allows *fluid and nonlinear development* of ideas, where agents can dynamically shift between proposals, revisit earlier concepts, and organically evolve the discussion. This leads to a more robust and well-rounded argument plan.

Additionally, current evaluation metrics for content diversity primarily measure lexical or semantic diversity, making them insufficient for assessing perspective diversity in long-form discourse. To address this, we develop a novel automatic metric specifically for evaluating perspective diversity. Our metric leverages extracting key ideas, comparing the uniqueness of each perspective, and aggregating these scores to determine overall perspective diversity. This approach assesses the variety of perspectives the model uses for argument writing.

We conduct experiments on argumentative essay writing using topics from the idebate and reddit/CMV portals, encompassing a wide range of domains. Both automatic and human evaluations indicate that our method produces outputs that are more diverse and coherent compared to those generated by baselines. Our key contributions are: (1) We propose a persona-based multiagent approach to ensure diverse perspectives in argument generation; (2) We develop a debate-driven planning that allows fluid and nonlinear development of ideas; (3) We design a novel metric for evaluating perspective diversity in long-form output.

2 Method

Given an input proposition (x) on a topic, our multi-agent framework generates an argument (y) with the following steps: (1) persona assignment, which creates and assigns an underlying persona to each agent; (2) debate-based planning, where agents collaboratively engage in debate and discussion to form a high-level plan; (3) argument writing that transforms the developed plan into a surface argument. The overall framework is shown in Figure 1.

2.1 Persona Assignment

Faced with a proposition on a controversial topic, people often form their opinions based on their underlying beliefs. This module generates and assigns a unique persona to each agent, representing their core beliefs. These personas serve as hidden variables that influence the agents' contributions during subsequent debate and writing tasks.

Persona Pool Creation. We instruct LLMs to create a pool of 5 to 10 personas, each embodying a distinct viewpoint relevant to the topic. We formalize a persona with a brief description and a claim on the topic, as illustrated in Figure 1. To ensure fairness and inclusivity, the model is directed to create personas representing a diverse range of communities and perspectives, which encourages the model consideration of nuance and in-group variation (Sorensen et al., 2024).

Persona Selection. After creating the persona pool, LLMs are prompted to select a combination of N personas from the pool and assign them to each participant, where N represents the number of participants. The model is guided to provide an explanation for each persona selection, ensuring that the chosen personas collectively contribute to a robust collaborative effort. We set N as 3 in our work.

2.2 Multi-agent Debate for Text Planning

Recent studies have highlighted the effectiveness of improving LLM performance with multi-agent collaboration (Li et al., 2023; Wang et al., 2023b; Du et al., 2023; Liang et al., 2023). We introduce a persona-based multi-agent debate for text planning, with each agent implemented as an LLM instance.

In our framework, the N agents form a *main team*, fostering collective discussions and developing a plan outlining the high-level logical flow. Additionally, we introduce a *critic agent* representing an opposing viewpoint. The role of the critic is to identify and challenge weaknesses in the main team's proposals. Incorporating such a critic is crucial, as a robust argument necessitates anticipating opposing perspectives and devising effective rebuttals during discussions.

After the agent initialization, the agents start a debate and express their opinions iteratively. This discourse continues until the main team members are satisfied and the critic agent is persuaded. Subsequently, the model synthesizes a final argument plan based on the discussion, representing the high-level logical flow of the text. Our debate-driven planning mirrors real-time discussions, wherein ideas evolve, face challenges, and undergo refinement in a nonlinear manner, ultimately resulting in a more cohesive and persuasive overall plan.

2.3 Argument Writing

The argument writing module then transforms the plan into a final argument. By employing the plan as high-level guidance, this module generates arguments in a controllable manner to ensure the output coherence. Our framework promotes thoughtful deliberation in the writing process by decomposing the text planning stage from end-to-end generation, enabling more polished and structured arguments.

3 Experiment Setup

Our experiments aim to evaluate the framework by exploring how the persona-based multi-agent system improves content diversity and enhances argument quality through debated-based planning.

3.1 Tasks

We evaluate our framework on argument essay writing (Bao et al., 2022). We collect 64 input propositions from `idebate.net` and reddit/CMV on various domains such as culture, politics, and education. Each proposition represents a controversial topic, like *"We should make all museums free of charge."* A model needs to generate a counter-argumentative essay to refute the proposition.

3.2 Model and Baselines

We implement all modules by prompting an LLM. For baselines, we include: (1) Directly prompting an LLM (LLM-E2E) to write an argument essay in an end-to-end manner; (2) Chain-of-Thought Prompting for content planning (LLM-Plan), where the model first generates an overall plan and then produce the argument (Wei et al., 2022); (3) AGENT-DEBATE: multi-agent debate for planning without persona assignment (Liang et al., 2023); (4) AMERICANO: decomposed argument generation with discourse-driven planning (Hu et al., 2023). We utilize ChatGPT as the backbone LLM for all methods. More details are in Appendix A.

3.3 Evaluations

Argument Quality. We employ both automatic and human evaluations. For automatic evaluations, we follow previous work and utilize GPT-based method to evaluate relevance and quality of the generated argument (Zheng et al., 2023; Chia et al., 2023). For human evaluation, we assess the persuasion and overall quality of generated arguments.

Semantic Diversity. To measure content diversity, we prompt models to generate 7 outputs for each input proposition. For semantic diversity, we first use self-BLEU (Zhu et al., 2018) to measure diversity among multiple generations for an input. As BLEU only captures word overlap, we also propose a self-Emb metric where we use embedding similarity to replace the BLEU score.

Perspective Diversity. We propose a novel perspective diversity metric, measuring how many unique perspectives the model generates to construct an argument. Concretely, for each input, a model generates M arguments $\{y_1, ..., y_M\}$. For a generated argument y_m, we first prompt ChatGPT to extract its main opinion points $O_m =$

	GPT Eval. (↑)		Diversity Eval. (↓)		
Method	**Rel.**	**Qual.**	**S-BLEU**	**S-Emb**	**Pers.**
LLM-E2E	3.56	3.63	24.54	87.66	73.77
LLM-Plan	3.59	3.48	20.54	86.54	73.37
AGENT-DEBATE	3.81	3.75	19.70	85.60	71.71
AMERICANO	3.75	**4.14**	22.93	85.59	72.85
Ours	**3.89**	3.91	**18.61**	**84.91**	**70.71**

Table 1: Automatic results. For GPT-based metric, we evaluate output quality (Qual.) and relevance (Rel.). For diversity, we report self-BLEU (S-BLEU), self-Emb (S-Emb) and Perspective Diversity (Pers.).

$\{o_{m1}, ..., o_{mn}\}$. Then for each opinion point o_{mi}, we compute its embedding similarity with the opinion points from all other $M-1$ arguments generated with the same input, and take the maximum similarity score (s_{mi}). The perspective diversity score of y_m is then computed as $s_m = \frac{1}{n}\sum_1^n s_{mi}$. The overall diversity score of the sample is the average of all arguments: $\frac{1}{M}\sum_1^M s_m$. A lower score indicates better perspective diversity achieved.

4 Result and Analysis

4.1 Main Results

The automatic results are shown in Table 1. For GPT-based evaluation, our model achieves the highest scores for output relevance and the second highest score for overall quality, demonstrating the effectiveness of our method. Notably, AGENT-DEBATE outperforms the directly prompted (LLM-E2E) and linear planning (LLM-Plan) baselines. This underscores the efficacy of leveraging multi-agent debate for text planning to enhance model performance in argument essay writing. However, it underperforms our model, indicating the advantage of our persona assignment for enhancing debate process. Additionally, in terms of diversity, our model significantly surpasses all baselines, producing outputs with both semantic diversity and rich perspectives. Conversely, LLM-E2E generates the least diverse outputs in terms of perspectives. This proves the effectiveness of persona assignment to enable the model to encompass a broader spectrum of viewpoints on subjective topics.

4.2 Human Evaluations

Due to the limitation of automatic evaluations, we also conduct human evaluations. Specifically, we randomly sample 30 inputs, and ask three human judges to evaluate the models outputs on aspects of persuasion and overall preference. [1] The results are shown in Table 2. Human judges consistently

[1] Details are provided in Appendix B.

Model	**Persuasion**	**Overall**
LLM-E2E	2.05 / 41.4%	1.99 / 36.8 %
LLM-Plan	2.21 / 48.3%	2.13 / 39.1 %
Ours	**2.31 / 51.7 %**	**2.47 / 66.7 %**

Table 2: Human evaluation results. The first position is the score, and the second position is the percentage of results ranked first (ties are allowed).

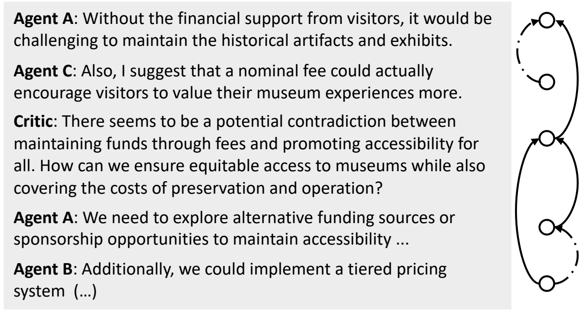

Figure 2: Snippet of the debate among agents for example in Figure 1. The right structure shows the logical flow, where solid arrow is oppose relation and dashed arrow is support.

rate our model outputs better than the baselines in both aspects. Particularly, our model generates outputs that cover a broader range of perspectives, thereby enhancing the overall persuasiveness of the argument. Moreover, our model is more frequently ranked as the top choice, further demonstrating its effectiveness in generating persuasive and high-quality argumentative essays.

4.3 Sample Analysis

In Figure 2, we show a snippet of debate process for the input *"We should make all museums free of charge"* as in Figure 1. The structure on the right illustrates the logical flow in a non-linear manner, showcasing how agents progressively discuss, revisit, and revise earlier points to address critics. By fostering an environment of ongoing dialogue and reflection with non-linear thinking, the internal multi-agent debate facilitates a more flexible and comprehensive planning process. We provide more sample outputs and additional discussions of weakness and future work in Appendix C.

5 Conclusion

In this study, we introduce a multi-agent debate framework with persona assignment for each agent to enrich perspective diversity and enhance persuasiveness in argument generation. Our debate-driven planning fosters fluid and nonlinear development of ideas for text planning, resulting in more robust and coherent argument plans. Experimental results across diverse topics demonstrate that our framework yields more diverse and superior arguments.

Limitations

Our work has several limitations that could be addressed in future studies. Firstly, effective argumentative essays often rely on supporting evidence to bolster claims. Humans typically seek out relevant knowledge or evidence to augment the persuasiveness of their arguments. Therefore, our framework could benefit from the integration of a knowledge retrieval module to incorporate external evidence. Secondly, while our focus in this work is on argument generation, our framework could be applied to other open-ended generation tasks, such as news article writing and story composition. Exploring these potential applications could strengthen the applications of our framework. Thirdly, our current persona creation module relies on LLMs to create a persona pool. People propose their opinions based on their values and norms, which we do not include in this process. Explicitly incorporating the norms and values would ensure the trustworthiness of the generated arguments, which can be incorporated in the future work.

Ethics Statement

Acknowledging the reliance of our framework on large language models, we recognize the possibility of generating fabricated and potentially harmful content due to inherent biases in the pre-training data drawn from heterogeneous web corpora for LLMs. Given the inability to fully control the language model generation process, there exists a risk of unintended biases persisting in the generated outputs. We strongly urge users to meticulously evaluate the ethical implications of the generated content and exercise prudence when employing the system in real-world contexts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. AEG: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. Planet: Dynamic content planning in autoregressive transformers for long-form text generation. *arXiv preprint arXiv:2203.09100*.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2023. Americano: Argument generation with discourse-driven decomposition and agent interaction. *arXiv preprint arXiv:2310.20352*.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

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

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express big five personality traits. *arXiv preprint arXiv:2305.02547*.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

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*.

Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici, and Fosca Giannotti. 2024. An overview of recent approaches to enable diversity in large language models through aligning with human perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 49–55, Torino, Italia. ELRA and ICCL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rose E Wang, Esin Durmus, Noah Goodman, and Tatsunori Hashimoto. 2022. Language modeling via stochastic processes. *arXiv preprint arXiv:2203.11370*.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2023a. Argument and counter-argument generation: a critical survey. In *International Conference on Applications of Natural Language to Information Systems*, pages 500–510. Springer.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Experimental Details

A.1 Dataset

In this work, we study zero-shot argumentative essay writing leveraging the large language models, and we select topics from `idebate.net` and reddict/CMV [2]. Each topic is a controversial proposition, such as *"We should make all museums free of charge."* We select 64 inputs covering different domain, and ensure they do not contain offensive contents. The model is asked to write a counter-argumentative essay to refute the proposition. The specific prompts we leveraged are presented from Figure 7 to Figure 9.

[2] `https://www.reddit.com/r/changemyview/`

A.2 Model Details

We set the number of agents for the main team as 3 in our experiments. We utilize ChatGPT as the backbone LLM, and use the *gpt-3.5-turbo-0301*[3] version. During inference, we set the temperature parameter as 1.0. During the planning process, we define the plan as a high-level outline that contains several main points, where each point can be supported by several sub-points. We also allow an optional acknowledgment point.

Baselines. (1) For LLM-E2E, we directly prompt an LLM to generate the output without explicit text planning. (2) For LLM-Plan, we first prompt an LLM to write a high-level plan, and the generate output based on the topic and plan. Similar to our model, we define plan as the same structure. This baseline is similar to the chain-of-thought prompting where the model first think about the high-level contents by generating the plan and then producing the final output. (3) AGENT-DEBATE is a multi-agent debate framework, which is a similar method to Du et al. (2023) where two agents debate on a topic. The original work utilize the multi-agents to enhance LLM reasoning ability. Differently, we leverage the debate for planning: we set one agent as the planner and the other as the critic, and the planner refines the idea through the debate process with the critic, subsequently producing a final plan for argumentative writing. (4) AMERICANO (Hu et al., 2023) is an argument generation framework that decomposes the generation based on argumentative discourse structures.

A.3 Automatic Evaluation Details

For GPT-based evaluation, we leverage the GPT4 model with the *"gpt-4o-2024-05-13"* variant. The prompt used for evaluating **relevance** (Chia et al., 2023) and **quality** (Zheng et al., 2023) are adopted from the original papers. For **diversity** evaluation, to compute the embedding diversity, we apply *"text-embedding-3-small"* model from OpenAI API to transform each output to an embedding, and then compute their cosine similarity.

For semantic diversity, besides self-BLEU, we design a self-Emb method, where we use cosine similarity between two output embeddings to replace the BLEU score. we apply *"text-embedding-3-small"* model from OpenAI API to transform each output to an embedding.

[3]https://platform.openai.com/docs/models

B Human Evaluation

For human evaluation, we hire three human judges to evaluate the model outputs on persuasiveness and overall preference. We ensure all judges are proficient English speakers with at least a Bachelor degree. We randomly select 30 inputs, and for each input we present the outputs of different models anonymously. All human annotators are graduate students based in US, and we pay them $12 per hour. We ask the human judges to rank the outputs based on each evaluation aspect, and ties are allowed. We then convert the ranks into scores by subtracting its ranking position from the total number of outputs (i.e., *model1* > *model2* > *model3* will lead to the score of 3 for model1, 2 for model2, and 1 for model3), as in Table 2.

For persuasiveness aspect, we ask the human judges to determine: whether the essay effectively challenges the initial proposition by providing convincing viewpoints from various perspectives with coherent logic; whether it is likely to persuade you to reconsider your initial position. For overall preference, we ask the human judge to evaluate on its overall quality and writing.

C Sample Outputs and Additional Analysis of Future Work

We first show different persona assignments for the same topic, as in Figure 3. The persona and claim represent the underlying themes of each perspective. As we can see, our model generates distinct viewpoints for the same topic, enhancing diversity in perspectives. Additionally, we present more comprehensive sample outputs generated by our models from Figure 4 to Figure 6. These examples illustrate how our debate-based planning effectively develops ideas and constructs logical plans aligned with agent personas, guiding the subsequent writing of final arguments. However, one potential improvement is that current arguments tend to focus on reasoning without sufficient evidential support. Persuasive arguments require factual evidence or expert opinions to strengthen their claims. Therefore, future work might focus on integrating explicit knowledge retrieval to enhance overall persuasiveness.

Topic: We should make all museums free of charge

Agents of main team 1

- Agent A - A museum employee: Making museums free would lead to budget cuts that could prevent museums from providing the quality of exhibits and educational programming.

- Agent B - An art collector: A free admission policy would lead to an influx of visitors who are not genuinely interested in the art, leading to more congestion, less space and consequently less enjoyment for art lovers.

- Agent C - A taxpayer: Free admission to museums would result in increased taxes, which would not only harm low-income individuals but would also impose an unnecessary burden on working and middle-class families who are already struggling to meet ends.

Agents of main team 2

- Agent A - A historian who values the preservation of artifacts: If museums are free, they may not generate enough revenue to properly maintain and protect historic artifacts and valuable works of art.

- Agent B - A community organizer advocating for equal access to cultural institutions: Charging admission fees to museums disproportionately affects low-income communities and excludes them from engaging in cultural experiences.

- Agent C - A museum curator trying to balance the budget: Without admission fees, museums may struggle to generate enough revenue to cover operational costs, resulting in reduced services and potentially closing the museum altogether.

Agents of main team 3

- Agent A - A museum administrator: Making museums free would result in the loss of crucial funding needed to maintain and improve exhibits.

- Agent B - A cultural studies scholar: Charging for museum admission is actually beneficial as it allows people to value the works more and fosters cultural appreciation.

- Agent C - A museum donor: Eliminating museum admission fees would deter future charitable donations and hamper the ability of museums to acquire new exhibits and fund construction projects.

Figure 3: Different Agent person assignment for the same topic.

Argument Plan:

① Acknowledgment:

- Recognize the value of free admission in promoting accessibility and attracting tourists.

② Financial Sustainability and Conservation

- Entrance fees are crucial for funding museum upkeep and conservation efforts.

- Alternative funding sources and sponsorships can supplement revenue without hindering accessibility.

③ Local Community Impact

- Implement a tiered pricing system to ensure locals have free or discounted access.

④ Visitor Engagement and Value

- A nominal fee can encourage visitors to engage more deeply with museum experiences.

- Thoughtful pricing strategies can enhance the overall value perception for visitors.

Argumentative Output:

While the notion of providing free admission to all museums may seem appealing at first glance, a deeper examination reveals that this proposition might not be as beneficial as it appears. Undoubtedly, there are advantages to offering free access, such as promoting inclusivity and attracting tourists. However, the financial sustainability and conservation of these cultural institutions must also be taken into account.

Entrance fees play a crucial role in funding the upkeep and conservation efforts of museums. Without this source of revenue, these institutions may struggle to maintain their collections, facilities, and educational programs. While it is true that museums can explore alternative funding sources and seek sponsorships, relying solely on these methods may not guarantee the financial stability needed to safeguard and preserve valuable cultural artifacts for future generations.

Moreover, considering the impact on the local community is essential. Implementing a tiered pricing system that offers free or discounted access to residents ensures that the local population can continue to benefit from their cultural heritage. By differentiating pricing for tourists, museums can strike a balance where visitors contribute more to support the institutions without excluding those who live in the area. This approach acknowledges the importance of local engagement and the preservation of community ties to these cultural landmarks.

In terms of visitor engagement and perceived value, a nominal fee can actually enhance the overall museum experience. When visitors invest financially, even in a small way, they are more likely to engage deeply with the exhibitions and educational resources available. Thoughtful pricing strategies can also create a sense of exclusivity and value, encouraging visitors to appreciate the cultural and historical significance of what they are experiencing.

In conclusion, while the idea of making all museums free of charge may seem well-intentioned, it is essential to consider the broader implications on financial sustainability, local community impact, and visitor engagement. By maintaining entrance fees and implementing strategic pricing models, museums can ensure their long-term viability while continuing to provide access to all, thus striking a balance between accessibility, conservation, and cultural value.

Figure 4: The full generated argument plan and output for input in Figure 1 and Figure 2

Topic: We should allow the use of electronic and internet voting in state-organised elections

- Agent A - Elderly voter advocate: Implementing electronic voting may disenfranchise elderly voters who are less familiar with technology and prefer traditional methods.

- Agent B - Privacy advocate: Electronic voting raises significant concerns about the privacy and confidentiality of voter information.

- Agent C - Skeptical traditionalist: Maintaining the traditional paper ballot system ensures transparency and trust in the electoral process, avoiding potential technological vulnerabilities.

Debate Process:

Agent A: I believe that implementing electronic voting may disenfranchise elderly voters who are less familiar with technology and prefer traditional methods. This can lead to a decrease in voter participation among the elderly population.

Agent B: Additionally, electronic voting raises significant concerns about the privacy and confidentiality of voter information. The security measures in place may not be foolproof, risking the integrity of the electoral process.

Agent C: Maintaining the traditional paper ballot system ensures transparency and trust in the electoral process. It avoids potential technological vulnerabilities that could be exploited to manipulate election results.

Critic: While the concerns raised are valid, electronic and internet voting can improve accessibility for voters with disabilities and those living far from polling stations. It also has the potential to increase voter turnout among the tech-savvy population.

Agent A: I acknowledge the accessibility benefits, but we must prioritize inclusivity for all voter demographics. Introducing electronic voting without addressing the concerns of the elderly population could create a significant divide in voter representation.

Agent B: Security should also be a top priority. Even with advancements in technology, the risk of cyber threats and hacking remains a serious issue. We cannot compromise the integrity of elections for the sake of convenience.

(...)

Argument Plan:

1. Implementing electronic and internet voting may disenfranchise elderly voters due to technological barriers:

    - Elderly voters may struggle to adapt to new technology, leading to decreased participation.

    - Preference for traditional methods should be respected to maintain inclusivity across all age groups.

2. Electronic voting raises concerns about privacy and confidentiality:

    - Security measures may not be completely effective, risking the exposure of voter information.

    - Safeguarding voter data is crucial to ensure the sanctity of the electoral process.

3. Maintaining the traditional paper ballot system ensures transparency and trust:

    - Paper ballots offer a tangible and easily verifiable voting method.

    - Avoiding technological vulnerabilities helps uphold the integrity of election outcomes.

Argumentative Output:

In the evolving landscape of modern technology, the proposition of implementing electronic and internet voting in state-organised elections has sparked debates among policymakers and citizens. While proponents argue for the convenience and efficiency that digital voting methods can bring, it is essential to critically consider the potential drawbacks and risks associated with such a significant shift in our electoral process.

One of the primary concerns with introducing electronic and internet voting is the potential disenfranchisement of elderly voters. This demographic group may face technological barriers that could hinder their ability to participate effectively in the electoral process. The unfamiliarity and complexity of new voting technologies could lead to decreased participation among elderly voters, ultimately undermining the core principle of inclusivity in our democratic system. Respect for the preferences of all age groups, including those who are more comfortable with traditional voting methods, is crucial to ensuring that every eligible voter has an equal opportunity to exercise their democratic right.

Furthermore, electronic voting raises legitimate concerns about privacy and confidentiality. The security measures put in place to protect voter information may not always be foolproof, leaving room for potential breaches and hacks that could compromise the integrity of the electoral process. Safeguarding voter data is paramount to upholding the sanctity of elections, and any vulnerabilities in the digital voting system could jeopardize the confidentiality of individual votes and the overall trust in the electoral outcome.

In contrast, maintaining the traditional paper ballot system offers a level of transparency and trust that electronic voting may not yet be able to match. Paper ballots provide a tangible and easily verifiable voting method, allowing voters to physically mark their choices and witness the deposit of their ballots into a secure ballot box. By avoiding the inherent technological vulnerabilities associated with electronic and internet voting, the integrity of election outcomes can be better protected, instilling confidence in the legitimacy of the results.

In conclusion, while the idea of electronic and internet voting may seem promising in terms of convenience and efficiency, the potential risks and drawbacks cannot be ignored. The disenfranchisement of elderly voters, concerns about privacy and confidentiality, and the importance of maintaining transparency and trust in the electoral process all warrant careful consideration before implementing such a transformative change. As we strive to modernize our electoral system, we must prioritize the security, inclusivity, and integrity of our elections to ensure a fair and reliable democratic process for all citizens.

Figure 5: A full output of our model generated argument.

Topic: No one over the age of 80 should be allowed to serve in government.

- Agent A - Elderly Rights Advocate: Age should not be a determining factor in someone's ability to serve in government as it goes against the principles of equality and non-discrimination.

- Agent B - Experienced Politician: Individuals over 80 often bring valuable experience and wisdom to government roles, making them capable and effective leaders.

- Agent C - Youth Activist: Limiting government service based on age overlooks the potential of younger generations to bring fresh ideas and perspectives to governance.

Debate Process:

Agent A: As an Elderly Rights Advocate, I strongly believe that age should not be a barrier to serving in government. Restricting individuals over 80 from government roles goes against principles of equality and non-discrimination.

Agent B: I, as an Experienced Politician, argue that older individuals bring valuable experience and wisdom to governance. Excluding them solely based on age can deprive the government of capable leaders.

Agent C: I, a Youth Activist, agree that age should not limit one's ability to serve. Younger generations can offer fresh perspectives and innovative ideas to government roles.

Critic: I support the proposition that no one over 80 should serve in government. Older age may bring health concerns and potential lack of energy or agility, affecting decision-making and efficiency in governance.

Agent A: We acknowledge the concerns about health and agility but believe that individual capabilities should be assessed on merit, not age alone.

(...)

Argument Plan:

1. Age Should Not Determine Government Service

   - Subpoint: Equality and Non-Discrimination Principles: Every individual should be evaluated based on merit and capabilities, not age.

   - Subpoint: Capability Is Timeless: Experience and wisdom are valuable assets in governance, regardless of age.

2. Embracing Diversity of Perspectives

   - Subpoint: Fresh Ideas from Younger Generations: Excluding older individuals limits the range of perspectives, hindering innovative governance.

   - Subpoint: Varied Experiences: Different age groups bring diverse life experiences that enrich decision-making processes.

3. Individual Assessment Over Arbitrary Age Limits

   - Subpoint: Merit-Based Evaluation: Evaluate individuals on relevant criteria like skills, experience, and abilities rather than imposing blanket age restrictions.

Argumentative Output:

In a world where age often dictates far too many aspects of life, the proposal that individuals over the age of 80 should be excluded from government service based solely on their age is not only discriminatory but also shortsighted. This notion fails to consider the fundamental principles of equality and non-discrimination that should underpin any fair and just society.

Every individual, regardless of age, should be evaluated based on their merit and capabilities. Experience and wisdom, often accumulated over decades, are valuable assets in governance. To dismiss these qualities based on a number is to overlook the timeless nature of capability. As society progresses, we must recognize that age should not be the sole determinant of an individual's ability to contribute to the governance of a nation.

Moreover, embracing diversity of perspectives is essential for effective governance. Excluding individuals over 80 limits the range of perspectives and experiences that can enrich decision-making processes. Younger generations may bring fresh ideas, but the wisdom and insight of older individuals are equally valuable in shaping policies that benefit all members of society. Varied experiences, derived from different age groups, contribute to a more holistic and inclusive approach to governance.

Instead of imposing arbitrary age limits, we should prioritize individual assessment based on merit. Merit-based evaluation considers relevant criteria such as skills, experience, and abilities, rather than imposing blanket age restrictions that fail to capture the nuanced capabilities of each individual. By evaluating candidates on their qualifications and aptitudes, we ensure that the most competent and capable individuals have the opportunity to contribute to government service.

In conclusion, age should not be a barrier to government service. Instead, we should recognize and value the diverse perspectives and experiences that individuals of all ages bring to the table. By embracing merit-based evaluations and rejecting arbitrary age restrictions, we can create a more inclusive and effective governance system that serves the needs of all members of society.

Figure 6: A full output of our model generated argument.

Prompt for persona pool creation:

Given a proposition: ##input_proposition
Background: You want to create a pool of 5 to 10 debate agents, who hold the opinions to refute the given proposition from different perspectives. Each agent should present a distinct viewpoint relevant to the proposition.
Task: Assign each agent a unique persona, described in one sentence, along with a corresponding claim that focuses on a specific perspective. Ensure that each agent provides a different viewpoint relevant to the proposition. To promote diversity and fairness, the agents should represent various communities and perspectives.
Please format your persona descriptions as follows, with each line being a json object:
{"agent_id": 0, "description": the_description_of_Agent0, "claim": the_claim_of_Agent0}

...

Prompt for persona selection:

Given a proposition: ##input_proposition
You need to build a team of three agents, to work together and collaboratively formulate a persuasive counterargument that refutes the given proposition. Now given the following candidates, where each candidate has a unique persona offering a different perspective relevant to the topic at hand. You need to select three agents that you think can together form a strong team to achieve the task. You also need to consider the diversity when selecting candidates. For each selection, give the reason why you select the candidate.

Candidate list:

###candidate_list

Please select three candidates and add a reason. Each line of output should be a json object as follows:

{"agent_id": 0, "description": the_description_of_Agent0, "claim": the_claim_of_Agent0, "reason": the_reason_of_selection}

...

Figure 7: Prompts for persona assignment.

Template for multi-agent debate

Background

Goal: Modeling a debate process to analyze a given proposition on a controversial topic, and formulate a well-structured counterargument plan to refute the proposition based on the debate discussion.

Additional Guidelines

- The discussion should be conducted for multiple rounds until the Main Team members are satisfied with their counterargument plan and Critic is persuaded.

- The discussion should provide a rigorous reasoning so that the logic flow is persuasive and coherent.

- Plan Quality: The plan should be abstract and concise. It should contain several main points, where each point can be supported by sub-points. There could be an optional acknowledgment point.

Main Team Agent

Participants and Roles

A Main Team of three members: Agent A, Agent B, and Agent C

- Stance: Oppose the proposition;

- Goal: Discuss together to propose a persuasive counterargument plan outlining the overall logical flow to refute the proposition.

- Specific Personas and claims of the team members:

- Agent A: persona_a;

- Agent B: persona_b;

- Agent C: persona_c;

Critic Agent

Participants and Roles: A Critic

- Stance: Support the proposition;

- Goal: You Disagree with the Main Team. Identify and challenge weaknesses in the Main Team's discussion, and debate with the Main team.

Figure 8: Prompts for multi-agent debate.

Prompt for surface argument writing:

Given a proposition: {proposition}

Write a persuasive and coherent counterargumentative essay to refute the proposition. You should transform the following plan into a coherent essay, which outlines the high-level logical flow of the counterargument.

- plan

{plan}
Note: ensure the essay is coherent and readable. You do not need to include section title from the plan.
- Counterargumentative essay:

Figure 9: Prompts for surface argument writing.