

Fact-driven Storytelling with Large Language Models

Xinyan Guo; Jiaqi Mo; Ying Hua Technical University of Munich, School of Computation Information and Technology, Research Group Social Computing

Motivation

- 1. How to generate argumentative messages in good quality?
- Use pyramid-structure for storytelling!
- How to generate a pyramid in good quality?
- LLM helps to generate arguments
- Finetuning BERT-based models for quality control!

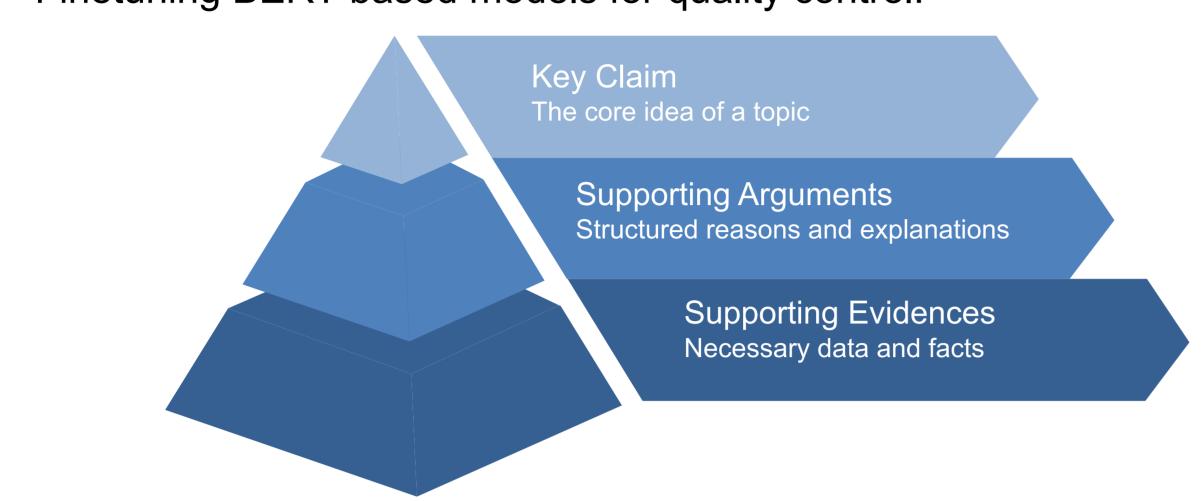


Figure 1. The Pyramid Principle for argumentative messages [2]

Evaluation

- Manually created pyramid-structured evaluation dataset including labels (e.g., "perfect" / "badclaim" / "badevidence") [5]
- Total 152 pyramids
- 87% accuracy

	Relevance	Support	Coherence	Overall
Accuracy	0.78	0.87	0.80	0.87
Precision	0.88	0.96	0.71	0.92
Recall	0.83	0.84	0.89	0.75
F1-Score	0.85	0.89	0.79	0.83

Table 1. Evaluation metric scores for different models

Methods

- Relevance: how well the arguments are directly related to and support the main claim
- Supportiveness: how effectively the evidences back up each argument
- Coherence: the logical flow and overall cohesiveness of the entire argument structure

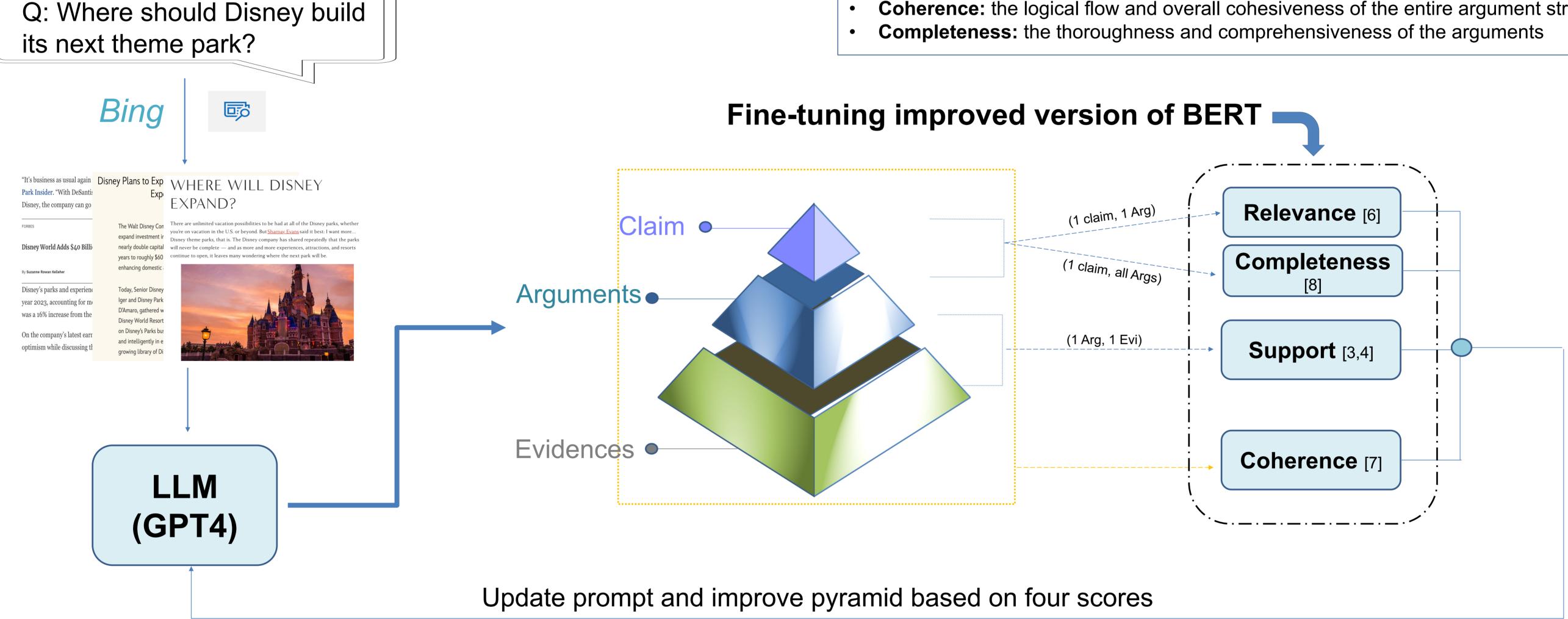


Figure 2. Pipeline of argument generation based on LLM

Results Claim Public transport should be free. Argument Argument Argument Free public transport would boost Free public transport would benefit Free public transport would have the economy. social benefits. the environment. Evidence Evidence Evidence Evidence Evidence Evidence According to a 2019 A 2017 report by the A 2018 study by the A 2020 study by the Research by the National Equity Atlas A 2018 study by the **Urban Institute found** European Environmen that free public Agency found that Association of Public cities with free public cities implementing transport increases free public transport Transport in 2019 transport reported free public transport workforce mobility systems saw a 15% indicated that free essential services higher overall quality experienced a 10% by 20%, enabling increase in consumer public transport led to such as healthcare, of life indicators, such decrease in CO2 more people to spending as residents a 25% reduction in education, and as reduced stress emissions due to saved money on traffic congestion in employment for lowlevels and increased access job reduced car usage transportation and income and leisure time, as opportunities and urban areas, resulting contributing redirected those residents spent less in lower pollution marginalized reducing significantly to savings towards other levels and improved communities, time and money on environmental unemployment local goods and commuting. air quality. promoting social equity sustainability. services. and inclusion. Figure 3. Example result pyramid generated from above process

Conclusions

- We proposed a machine learning approach to optimize the quality of a pyramid-structured argumentative message based on LLM.
- We developed a quality measure system by finetuning BERT-based models and calculating 4 different scores for each input pyramid.
- We tested the performance of our models using a manually created pyramid-style dataset and achieved an overall accuracy of 87%.

Check our code and data on GitHub!



Contact

Xinyan Guo, Jiaqi Mo, Ying Hua Technical University of Munich

Email: xinyan.guo@tum.de, ge74toz@mytum.de, ying.hua@tum.de

References

- 1. Chen, G., Cheng, L., Tuan, L. A., & Bing, L. (2023). Exploring the potential of large language models in computational argumentation. arXiv
- preprint arXiv:2311.09022. 2. Minto, B. (2009). The pyramid principle: logic in writing and thinking. Pearson Education.
- Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., & Slonim, N. (2019). Are you convinced? choosing the more convincing evidence with a Siamese network. arXiv preprint arXiv:1907.08971.
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations.
 - In International conference on machine learning (pp. 1597-1607). PMLR. https://www.kialo.com/
- Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., & Slonim, N. (2020, April). A large-scale dataset for argument quality ranking: Construction and analysis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 7805-7813). 7. Zhu, D., Wu, W., Song, Y., Zhu, F., Cao, Z., & Li, S. (2024). CoUDA: Coherence Evaluation via Unified Data Augmentation. arXiv preprint
- arXiv:2404.00681. 8. Stab, C., & Gurevych, I. (2017, April). Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th* Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (pp. 980-990).