Fact driven Storytelling with Large Language Models

Xinyan Guo
Technical University of Munich
xinyan.guo@tum.de

Jiaqi Mo
Technical University of Munich
ge74toz@mytum.de

Ying Hua
Technical University of Munich
ying.hua@tum.de

Abstract

A new approach is proposed for generating argument pyramids using Large Language Models (LLMs), combining the Bing API and a fine-tuned Flan-T5-Large model to efficiently extract accurate evidence. By pre-processing the evidence, this approach effectively reduces the hallucination problem common to the direct application of LLMs and ensures the reliability of argument generation. Our structured approach improves the relevance, support, coherence, and completeness of arguments and evidences, thereby enhancing the persuasiveness of the content. In addition, we develop customized BERT-based metrics to rigorously assess these dimensions. This scalable solution is particularly well suited for decision-making processes in consulting and academic research, especially in scenarios where high-quality argument content is required.

1 Introduction

As artificial intelligence continues to rapidly evolve, strong tools of organizing complex information into coherent and persuasive arguments are especially important so as to better inform decision-making processes across different fields including law, business and education. Minto (1981) makes the case in her book, the Pyramid Principle is introduced and widely used for its effectiveness in organizing thoughts in a clear, logical manner. This principle underlines the significant takeaways summarized at a higher level, supported with detailed layers to form a pyramid. In this way, the persuasiveness of arguments is increased while not losing any information.

However, manual application of the Pyramid Principle can be both time-consuming and subject to human error, limiting its scalability and efficiency in professional settings. In this context, we propose our project to address this gap by adapting a machine learning based model using Large Language Models (LLMs) for generating well organized arguments automatically following the Pyramid Principle. Combining AI-based methods to this integration will change the way we process and interpret information, providing superior precision and transparency.

To ensure the reliability and effectiveness of the generated arguments, we employed BERT-based models, which have become very popular as a base model for fine-tuning on downstream NLP tasks. These models are further fine-tuned to assess the quality of arguments, which should conform ideally specific standards of relevance, supportiveness and coherence. The inspiration for combining LLMs for generating content and BERT-based models for evaluation came from Trinh et al. (2024). In this study, researchers used language models to further explore and generate solutions to geometry problems by DeepMind's AlphaGeometry model, demonstrating how language models can assist in solving complex problem-solving tasks, guiding the methodology of our project.

Our study seeks to enhance discussion quality, and thereby decision-making not only by simplifying communication procedures but also providing a tool that can be re-used across multiple professional domains. Based on the Pyramid Principle, our project automates the structuring of information to enable a more efficient and effective data-based communication.

2 Related Work

The evolution of Large Language Models (LLMs) has significantly impacted various domains, with creative writing and story generation emerging as particularly transformative areas. In this section, we will discuss the latest research surrounding the development and evaluation of LLMs in storytelling.

2.1 Story Generation using LLM

The application of LLMs extends beyond traditional NLP tasks, with studies exploring their use in generating definitions for financial terminology, extracting knowledge entities in astrophysics articles, and aiding in neurosymbolic robot action planning (Jhirad et al., 2023; Shao et al., 2024). These applications highlight the adaptability and potential of LLMs in diverse fields, showcasing their ability to revolutionize processes and enhance decision-making. Therefore, in our study, we are going to generate argumentative messages using LLMs.

The field of automatic story generation has advanced significantly with the introduction of Large Language Models (LLMs), which have minimized the necessity for extensive human involvement in creating stories (Valentini et al., 2023). Recent progress has concentrated on improving the quality of generated stories by integrating elements such as plots and commonsense knowledge (Xie et al., 2023). LLMs have proven their ability to produce stories that are not only coherent but also demonstrate logical consistency and global coherence, surpassing previous models (Guan et al., 2020).

Furthermore, LLMs have been applied in various research initiatives to enhance different aspects of story generation. For example, they have been utilized effectively in generating story analogies, underscoring their capacity to create high-quality analogical content (Jiayang et al., 2023). Additionally, LLMs have exhibited potential in elevating user story quality, indicating their versatility for diverse applications across various industries (Zhang et al., 2024).

2.2 Evaluation and better Storytelling

To evaluate argument quality, researchers have explored various methodologies utilizing advanced language models. Favreau et al. (2022) demonstrated the effectiveness of LLMs in recognizing arguments but highlighted limitations in quantifying argument quality compared to contrastive learning methods. These studies emphasize the complexity of evaluating argument quality and the necessity for nuanced approaches.

Furthermore, Zhang et al. (2023) emphasized the potential of LLMs, particularly instruction-tuned models, as effective automatic dialogue evaluators, showcasing their robust language understanding capabilities. This suggests that LLMs can be pivotal in assessing argument quality by leveraging their language processing abilities.

Other researches defined some conceptual notions of what is actually meant by argument quality, considering either the notions of maximal or minimal quality (Wachsmuth et al., 2024). The minimal quality measures an argument's evaluability. Inspired from this, we evaluate our generated arguments from different minimal aspects, such as coherence, relevance and support, in the later Section 4.

3 Pipeline Architecture for Generating Argument Pyramids with LLMs

In our system, we utilize a sophisticated pipeline that leverages Large Language Models (LLMs) to generate well-structured argument pyramids. The pipeline, as illustrated in Figure 1, integrates multiple components to ensure that the generated arguments are not only relevant and supportive of the main claim but also coherent and complete. This section outlines the key stages of our pipeline, demonstrating the workflow from query initiation to the output of structured argument pyramids.

3.1 Initial Query and Data Retrieval

The process begins with a query input, exemplified by the question, "Where should Disney build its next theme park?" To gather pertinent information, we first utilize the Bing API to search for relevant data across a vast array of sources. This initial retrieval step ensures that the data fed into the system is current and relevant, providing a solid foundation for the subsequent generation of arguments.

3.2 Evidence Extraction Using Flan-T5-Large

Upon retrieving relevant documents, the pipeline employs the Flan-T5-Large model, which is responsible for extracting key pieces of evidence from the fetched documents. This model is adept at distilling the essential facts and figures that are most pertinent to the query, ensuring that the information used to support the arguments is precise and pertinent.

3.3 Argument Generation with LLM (GPT-4)

Following evidence extraction, the Large Language Model, specifically GPT-4, takes center stage to construct arguments based on the extracted evidence. This stage is crucial as GPT-4 synthesizes the information into a coherent argument structure,

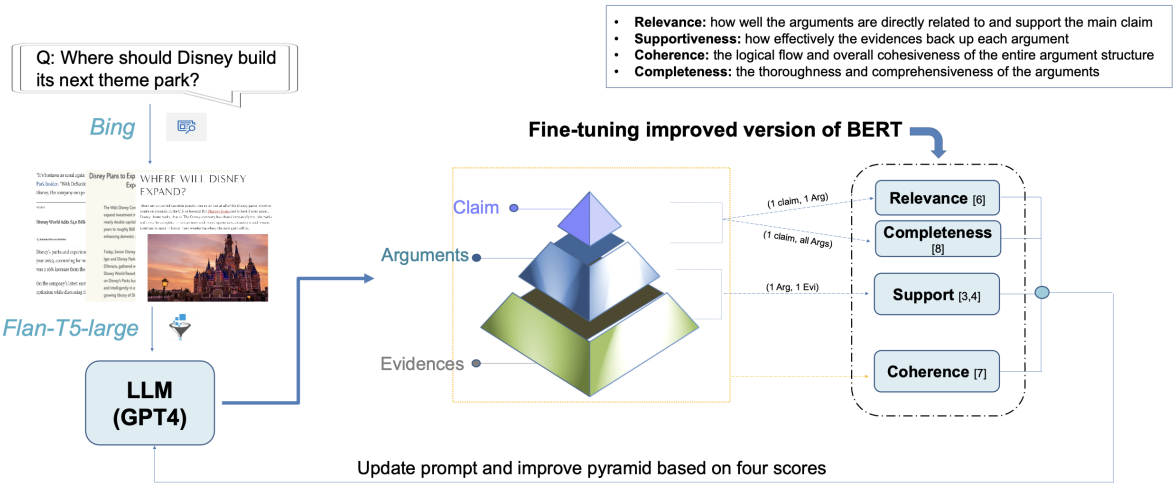

Figure 1: Pipeline of argument generation based on LLM

aligning with the Pyramid Principle, which advocates for a clear, logical progression of ideas.

3.4 Evaluation Using BERT

To refine the outputs further, an enhanced version of BERT, fine-tuned for our specific needs, evaluates the arguments based on four critical criteria: relevance, support, coherence, and completeness. This evaluation is crucial to ensure that each argument not only supports the main claim effectively but also integrates well with other arguments to form a logically consistent and comprehensive argument pyramid.

3.5 Update Prompt and Improve Pyramid

When scores for any component fall below a predefined threshold, determined through analysis of a manually constructed test dataset, the system automatically generates new prompts directing the LLMs to enhance the specific underperforming parts. This tailored approach helps fine-tune the arguments, focusing on areas that need improvement to enhance clarity and effectiveness. Through this methodical refinement process, each section of the argument pyramid is optimized, ensuring the overall quality is consistently high.

4 Model Design and Methodology

4.1 Relevance

In assessing the relevance of an argument to a particular claim, it is crucial to examine the extent to which the argument supports the claim. This is particularly important in areas such as debate, consulting, and public policy analysis. To improve the way we assess the relevance of arguments, we developed a model based on cutting-edge machine learning techniques and optimally trained it using data from the IBM/Argument-Quality-Ranking30k dataset (Gretz et al., 2020). This dataset contains 30,497 arguments around 71 controversial topics, carefully curated to train models that can focus on argument quality and relevance. We collected these arguments through a combination of crowdsourcing and expert annotation, ensuring a rich and diverse collection of viewpoints, each associated with a topic it supports or opposes.

Relevance in this context is defined as the degree to which an argument substantively supports or contests a claim within the context of a given topic. This was assessed by the likelihood of an argument being used effectively in a speech or debate on the topic, as judged by the annotators. The MACE-P scoring function (Habernal and Gurevych, 2016) was used as a key metric for evaluating the quality of annotations. MACE-P is an unsupervised item-response model that calculates the probability of each annotation being correct, adjusting for annotator reliability. This scoring function was essential for deriving continuous quality labels from binary annotations, helping to refine the assessment of argument relevance. The relevance model was fine-tuned using BERT's pre-trained model (Devlin, 2018) to accommodate the nuances of argument relevance assessment.

The fine-tuning was adapted to a regression task, aimed at predicting a relevance score between claim and argument. Adaptations included modifying BERT's typical classification output layer to a sigmoid function, allowing for the output of continuous values representing the relevance scores

between 0 and 1. Additionally, the standard cross-entropy loss was replaced with mean squared error, focusing on the discrepancies between the predicted relevance scores and the actual labels derived from the MACE-P scoring. The fine-tuned relevance model establishes a strong framework for quantitatively evaluating the relevance of arguments to claims. This model not only draws on the detailed annotations provided in the IBM-Rank-30k dataset but also takes full advantage of BERT's advanced encoding capabilities, ensuring both accuracy and reliability in the assessment process.

4.2 Support

The support model addresses the relation between arguments and corresponding evidences. The degree to which the evidence supports the argument is represented by a support score generated by the support model. As compared to predicting a label about whether the evidence supports the argument, this support score provides a quantitative measure, which contributes to further refining how well the evidences (including common sense, facts, or studies) supports the arguments.

Our training approach for the support model is inspired by SimCLR (Chen et al., 2020). It was initially applied in the field of learning visual representations and was found to outperform other methods for self-supervised and semi-supervised learning. Since there are many commonalities between visual and textual embeddings, we propose a contrastive learning approach based on this to train the support model. The framework of the support model is composed of: 1. Input data pairs including arguments and corresponding positive and negative evidences. 2. A RoBERTa encoder for fine-tuning based on contrastive learning. 3. A module that uses the embedding generated by the fine-tuned RoBERTa encoder to compute support score.

While SimCLR uses data augmentation operations to generate positive and negative pairs, we collect a training dataset containing positive and negative (argument, evidence) pairs from the existing dataset. Our dataset is extracted from IBM-EviConv (Gleize et al., 2019). It consists of arguments, positive evidences, and negative evidences. Each argument corresponds to one positive evidence and five negative evidences. As there are more than five negative evidences aligned with the corresponding argument and positive evidence, the top 5 are selected according to their convincingness scores.

Based on the above dataset, we fine-tune the pre-trained RoBERTa encoder to obtain sentence representations for each pair of argument, positive evidence, and negative evidences, respectively. During training, we freeze most of the layers of the pre-trained RoBERTa encoder, leaving only the last three layers unfrozen in order to update the weights. The sentence representations are taken from the output of the [CLS] tokens at the last layer, which is reported to aggregate information from the entire sentence (Devlin et al., 2019). At the top of the RoBERTa encoder, we add a layer to compute the contrastive loss, which aims to maximize the similarity between embeddings of positive pairs and minimize it of negative pairs. We adopt the normalized temperature-scaled cross-entropy loss (NT-Xent loss) from Chen et al. (2020) as our contrastive loss. In our setting, the positive pair contains the argument and its corresponding positive evidence from the dataset, while the negative pair consists of the same argument and its negative evidences.

After training, the RoBERTa encoder adapted by our dataset could generate sentence representations that capture the important features of the sentence to determine the degree of support. In view of this, we add a module that outputs a support score, which is obtained by computing the cosine similarity between an argument and an evidence embedding generated by the fine-tuned RoBERTa encoder. The support score (range: [-1, 1]) indicates the degree to which the evidence supports the argument, the closer to 1 the more supportive it is.

4.3 Coherence

In the realm of debate and discussion, coherence measures the logical consistency of arguments in supporting a central claim. This logical consistency is key to the persuasiveness and clarity of arguments across various platforms such as formal debates, academic discourse, and policy development. A coherent argument ensures that every component logically and seamlessly supports the main thesis without any contradictions or irrelevant diversions. Therefore, developing a precise measure of coherence entails evaluating the strength and flow of connections that bind the claim and its supporting arguments.

Presented at NAACL 2024, the COUDA (Coherence Evaluation via Unified Data Augmentation)

(Zhu et al., 2024)model introduces a cutting-edge method for assessing argument coherence through innovative data augmentation techniques. This model tests the resilience and logical consistency of argumentative texts by intentionally disrupting their coherence. On a global level, COUDA rearranges sentences, challenging the overarching logical flow and generating examples of global incoherence. Locally, it produces sentences that do not fit contextually with surrounding text, providing a richer and more complex array of negative samples compared to traditional rule-based methods. This approach leverages post-pretraining of language models along with finely tuned difficulty controls to craft these diverse samples.

The COUDA model utilizes ALBERT-xxlarge (Lan et al., 2019), selected for its proven ability to capture sentence-level coherence during tasks like sentence order prediction. This ensures that the model can precisely track the flow of information between sentences, crucial for effective coherence assessment. Comprising 233 million parameters, the model is meticulously fine-tuned using a balanced dataset of both coherent (positive) and incoherent (negative) samples. The training process is optimized with a batch size of 32 and a learning rate of 1e-5, achieving convergence within about 3,000 steps.

In our project, we've adapted the COUDA model to suit the unique requirements of evaluating argument pyramids, where traditional models of local coherence assessment might not suffice. In argument pyramids, adjacent sentences do not necessarily belong to the same line of reasoning as they could be part of different sub-arguments or evidential supports. Therefore, we adjusted COUDA to consider each "mini-pyramid" – a single argument and its corresponding evidences – as an individual unit for local coherence evaluation. This adaptation allows us to assess coherence more appropriately across the broader structure of the entire argument pyramid, referred to as global coherence. By doing so, we ensure that our coherence evaluations are not only more precise but also more reflective of the actual logical structure and effectiveness of the arguments within their specific contexts.

4.4 Completeness

To increase the credibility of a given claim to persuade the audience, it is important to explain the claim in a comprehensive and complete manner. In other words, the arguments extended from this claim should provide sufficient reasons for accepting the claim. Thus, a complete set of arguments is able to account for all aspects involved in the claim and draw logical reasoning. We aim to train a model to assess the completeness of arguments with respect to the claim, ensuring that the arguments generated by GPT-4 are sufficient to support the claim.

We utilize an existing dataset from Stab and Gurevych (2017) as our training dataset. It consists of 1029 labeled body paragraphs extracted from a corpus of 402 student essays. Each body paragraph is labeled as either sufficient or insufficient, indicating whether the claim (the starting sentence) is supported sufficiently by all arguments in this paragraph (the remaining sentences). Moreover, we remove the connectives in the paragraph to prevent their impact on determining the completeness of arguments. Based on this dataset, we fine-tune the pre-trained DeBERTa model (He et al., 2020) for a binary classification task. This pre-trained model is chosen for its strong performance with fewer data in classification tasks. The training results show that the fine-tuned model performs remarkably well in terms of accuracy, precision, recall, and F1 score. Finally, we connect the fine-tuned model with a softmax, and treat the probability of being classified as sufficient arguments to be the completeness score.

However, since the dataset is limited to the tendency to label arguments containing a single sample as insufficient, it lacks a comprehensive view of completeness. To fill this gap, we propose an additional approach inspired by Pan et al. (2023), where multi-step reasoning is utilized to verify the claim given evidence. Nevertheless, in our case, the claims are standpoints rather than facts that can be directly verified as in the paper. Therefore, we propose to utilize GPT-4 to generate questions around the claim to check the completeness of arguments. These questions should take into account several aspects related to completeness, such as social and ethical factors. Following this, we further utilize GPT-4 to assess the completeness of arguments according to the extent to which the arguments address these questions. For each generated question we define a rating scale with the range [0, 1] and a step size of 0.25, indicating the degree to which the arguments address this question. Finally, we average the scores of all questions as the completeness

score.

Nevertheless, the approach based on GPT-4 is also restricted to the uncontrollability and quality of the generated questions and the non-transparent processing of GPT-4. Thus, we combine the result of the classification and GPT-based approach to obtain the final completeness score to balance the limitations.

4.5 Evidences extraction

In our project, efficiently extracting useful evidence from articles presents an intriguing challenge. To tackle this, we've implemented the fine-tuned Flan-T5-Large Model, inspired by the paper (Chen et al., 2023). This paper explores dense retrieval methods, crucial for obtaining relevant context or world knowledge in open-domain NLP tasks. It highlights how the choice of retrieval unit—whether a document, passage, or sentence—significantly impacts the performance of retrieval and downstream tasks.

Distinct from traditional methods that use passages or sentences, the paper have been introduced a novel retrieval unit called a "proposition" for dense retrieval. Propositions are defined as atomic expressions within the text, encapsulating a distinct factoid in a concise, self-contained natural language format. We conducted empirical comparisons across different levels of retrieval granularity and found that proposition-based retrieval markedly outperforms traditional passage or sentence-based methods. Moreover, retrieving by propositions enhances downstream QA tasks' performance, as the texts retrieved are more focused and packed with question-relevant information. This reduces the need for processing lengthy input tokens and minimizes the inclusion of irrelevant details.

The primary purpose of our model is to extract and generate atomic propositions from documents, enabling more effective and token-efficient retrieval of evidence. By streamlining the evidence extraction process, we significantly reduce the number of input tokens required, thereby cutting down on the computational overhead associated with using APIs like GPT. This approach not only speeds up information processing but also improves the cost-effectiveness of the entire retrieval and processing workflow.

5 Evaluation

The BERT-based models in Section 4 assess key quality dimensions of the pyramids, with respect to relevance, support, coherence, and completeness. To determine how well these models perform, we must find a method of evaluation. In this section, we will discuss about the evaluation procedure.

To evaluate these models, we manually curated a dataset of 152 pyramids with varying levels of structural integrity. The topics of pyramids are gathered online from different websites, including debate websites, and we revised the level of quality with the help of LLM. These pyramids were annotated with the following labels:

- **Perfect**: A well-structured pyramid where the claim, arguments, and evidence are coherent, relevant, and logically connected. This serves as the positive example in our dataset.
- **Bad Claim**: The claim is not relevant to the arguments and evidence, though the supportiveness of the evidence toward the arguments remains unchanged.
- **Bad Argument**: Similar to Bad Claim, this label indicates that the arguments are not relevant to the claim, though the claim and evidence may still be related.
- **Bad Evidence**: The evidence does not support the arguments anymore, but the claim and arguments remain relevant and coherent with each other.

This annotated dataset was utilized to test and refine the performance of the BERT-based models across multiple quality dimensions.

To quantify the performance of our models, we employed standard evaluation metrics:

- **Accuracy** measures the proportion of correct predictions (both true positives and true negatives) made by the model relative to the total number of predictions.
- **Precision** (also known as Positive Predictive Value) reflects the model's ability to correctly identify positive instances, i.e., how many of the predicted positive cases were actually correct. High precision is important when the cost of false positives is high.

- **Recall** (also known as Sensitivity or True Positive Rate) captures the model's ability to correctly identify all relevant positive cases, i.e., how many of the actual positive cases were correctly predicted by the model. High recall is essential when the cost of false negatives is high.

- **F1-Score** is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between precision and recall, particularly useful when the class distribution is imbalanced.

	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 scores for different models

The results are summarized in Table 1, with the final model achieving an overall accuracy of **87%**. The precision and recall metrics highlight the model's ability to correctly identify well-structured pyramids while effectively handling sub-optimal cases. For instance, our model demonstrated strong performance in assessing **Support** with a precision of **0.96** and **Relevance** with a recall of **0.83**, indicating its robust capacity in these areas.

This evaluation is critical in understanding the efficacy of using BERT-based models to assess complex argumentative structures. Our findings suggest that while these models perform well across multiple dimensions, there is room for improvement in assessing **Coherence**, as indicated by relatively lower precision and recall scores in this area. **Completeness** is not considered here, as it is reletively hard to define a clear criteria of completeness and create those pyramids manually.

Future work could involve expanding the dataset, further refining the model architecture, or experimenting with ensemble approaches to improve these specific metrics. This evaluation serves as a foundational step toward more automated and fine-grained assessment of argumentative text structures, with potential implications for educational tools, automated debate scoring, and content generation systems.

6 Results

For instance, consider the question: "Where should Disney build its next theme park?" After utilizing our generator tool, we arrive at an Argument Pyramid, as illustrated in the appendix.

7 Limitation

1. **Hallucination of LLMs:** While generating argument pyramids based on factual evidence with LLMs effectively reduces hallucination issues, it still cannot entirely prevent the production of content that is inconsistent with the facts.

2. **Dimensional Incompleteness:** The four dimensions of relevance, supportiveness, coherence, and completeness do not fully represent the quality of an argument pyramid; other dimensions such as appropriateness and deliberation need to be measured.

3. **Model Alignment:** The four different BERT-based models used in our metrics also have some model alignment issues, as the datasets chosen for fine-tuning do not perfectly match our intentions, and the models do not always directly assess these dimensions.

4. **Test Dataset Limitations:** Our test dataset is manually constructed, and the human annotations involve biases. The small size of the dataset might lead to domain shift issues, causing our test results to be inaccurate.

5. **Feedback and Iterative Improvement:** During iterative improvement, merely changing the prompt sometimes fails to have an effect and can instead lower the scores for that section.

6. **Cost and Efficiency:** Running GPT-4 on a T4 GPU for six minutes costs about $2 to produce results, which is expensive and time-consuming. However, if using a 3.5 turbo or 4o, because we use regular expressions to extract arguments and evidence from the LLM's responses, it sometimes fails to capture them.

7. **Alignment Complexity:** Overall, our project involves a complex alignment issue among three parties: the user, the LLM, and the four models in the metrics.

8 Conclusion

In this project, we have successfully developed a robust methodology for generating argument pyramids using Large Language Models (LLMs) that significantly enhances the structure of persuasive arguments. By integrating Bing API, finetuned Flan-T5-Large, and customized BERT models, we effectively address common illusions in the output of LLMs, thus improving the relevance, support, consistency, and completeness of arguments. This system not only simplifies the process of creating argument content, but also ensures high-quality output suitable for consulting and academic research. Although challenges remain in model calibration, test dataset limitations, and computational efficiency, our work provides a solid foundation for future improvements. Next, we plan to further optimize the dataset used for fine-tuning, extend the evaluation criteria, and employ techniques such as RLHF (Reinforcement Learning with Human Feedback) to better align the goals of users, LLMs, and evaluation metrics. These measures aim to overcome existing limitations and improve the accuracy and reliability of our argument generation.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

C. Favreau, A. Zouaq, and S. Bhatnagar. 2022. Learning to rank with bert for argument quality evaluation. *The International FLAIRS Conference Proceedings*, 35.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a siamese network. *arXiv preprint arXiv:1907.08971*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

J. Jhirad, E. Marrese-Taylor, and Y. Matsuo. 2023. Evaluating large language models' understanding of financial terminology via definition modeling. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacifi.*

C. Jiayang, L. Qiu, T. Chan, T. Fang, W. Wang, C. Chan, D. Ru, Q. Guo, H. Zhang, Y. Song, Y. Zhang, and Z. Zhang. 2023. Storyanalogy: deriving story-level analogies from large language models to unlock analogical understanding. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Barbara Minto. 1981. The pyramid principle: Logic in writing and thinking.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

W. Shao, R. Zhang, P. Ji, D. Fan, Y. Hu, X. Yan, C. Cui, Y. Tao, L. Mi, and L. Chen. 2024. Astronomical knowledge entity extraction in astrophysics journal articles via large language models. *Research in Astronomy and Astrophysics*, 24:065012.

Christian Stab and Iryna Gurevych. 2017. 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*, pages 980–990.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625:476 – 482.

M. Valentini, J. Weber, J. Salcido, T. Wright, E. Colunga, and K. von der Wense. 2023. On the automatic generation and simplification of children's stories. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. *ArXiv*, abs/2403.16084.

Z. Xie, T. Cohn, and J. H. Lau. 2023. Can very large pretrained language models learn storytelling with a few examples?

Chen Zhang, L. F. D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2023. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *AAAI Conference on Artificial Intelligence*.

Z. Zhang, M. Rayhan, T. Herda, M. Goisauf, and P. Abrahamsson. 2024. Llm-based agents for automating the enhancement of user story quality: an early report. *Lecture Notes in Business Information Processing*, pages 117–126.

Dawei Zhu, Wenhao Wu, Yifan Song, Fangwei Zhu, Ziqiang Cao, and Sujian Li. 2024. Couda: Coherence evaluation via unified data augmentation. *arXiv preprint arXiv:2404.00681*.

A Example Appendix

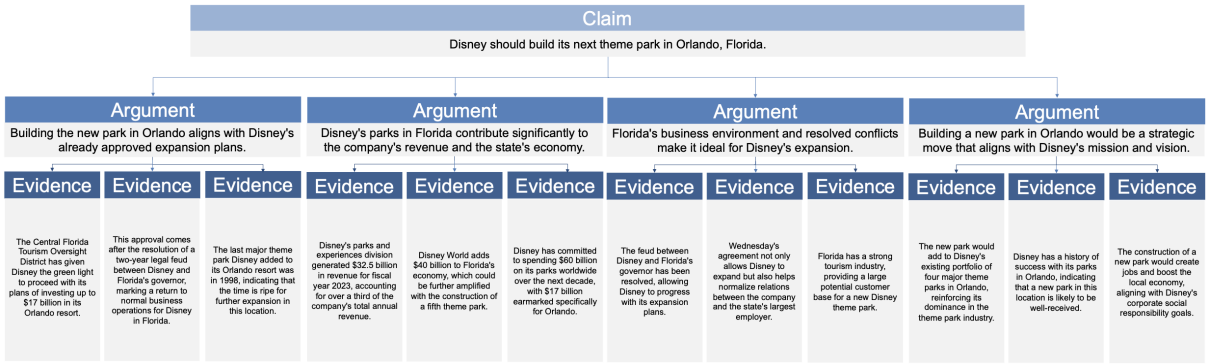

Figure 2: Argument pyramid example