Enhancing Argument Validity and Novelty Prediction: A Combined Framework of Feature-Based and Fine-Tuned GPT Models

Sean Zhou and **Owen Lin** and **Zhengxing Cheng**
seanzhou1207@berkeley.edu
shaochen_lin@berkeley.edu
zhengxing_cheng@berkeley.edu

Abstract

We predict the validity and novelty of a conclusion for a given premise on a defined topic, a subtask in the field of Argument Mining. This report describes our experiments in feature-based and fine-tuned GPT-based models. Our methodology incorporates topic data and used a domain-specific SBERT model in natural language inference to extract features. We also employ knowledge graphs to obtain semantic relationships between premise-conclusion entities. Integrating knowledge graph features and SBERT embeddings provides the highest performance among all tested combinations. Furthermore, We investigate fine-tuned GPT like Instruction GPT and implement a novel recurrent boosting fine-tuned framework called Recurrent GPT with the prompting chain. The best model, which combines Recurrent GPT and Instruction GPT, shows significant improvement over the RoBERTa baseline and achieves state-of-the-art performance among existing models.

1 Introduction

Fueled by the growing prevalence of online discussions and user-generated content, argument mining has generated significant attention. This branch of natural language processing is focused on extracting, analyzing, and identifying relationships among argument units (Lawrence and Reed, 2019), providing structured analysis of diverse opinions and reasoning within debates, reviews, and social media.

Specifically, the shared task of ACL's 9th Workshop on Argument Mining (2022) puts an emphasis on the validity and novelty of a premise-conclusion pair. Validity is defined as the degree a conclusion is justified by its premise. Conclusions that restate or summarize the premise are also considered valid and thus are typically not novel. Conversely, novelty is defined as the extent of new content in the conclusion beyond what the premise provides. A novel conclusion should still be related to the topic but contain segments that serve as a logical extension to the premise (Heinisch et al., 2022).

Previous research demonstrates that creating a conclusion that satisfies both validity and novelty is difficult, indicating a trade-off between successfully meeting the two criteria (Heinisch et al., 2022). In our study, we tackle Task A of the ArgMining 2022 challenge, which aims to predict an argument's validity and novelty based on the related topic and premise. We explore various combinations of NLP features, encompassing domain-specific SBERT and knowledge graphs in the feature-based model. Additionally, we address the task of validity and novelty prediction using an innovative fine-tuned GPT framework that leverages Recurrent GPT and instruction GPT techniques, ultimately achieving state-of-the-art performance.

2 Literature Review

As the arguments usually revolve around real-world events and cover a wide range of subjects, knowledge graphs are essential components in many prior argument-mining approaches. Saadat-Yazdi et al. used the WikiData knowledge graph (Vrandečić and Krötzsch, 2014) to extract semantic relationships between premises and conclusions. They identified the shortest paths between premise and conclusion entities and built numerical features like **Irrelevancy** and **Average Distance**. ConceptNet, a semantic network providing common-sense knowledge to machines, offers an alternative (Singh, 2002). Team ACCEPT achieved the best performance in novelty classification using ConceptNet's path construction and semantic relatedness features (Speer et al., 2017). In Figure 3, we illustrate a premise-conclusion pair related to the topic *Vegetarianism*, highlighting distinctions between entities with short or long intermediary paths as well as disconnections.

Another approach frequently utilized is the Generative Pre-trained Transformer (GPT) model. In previous tasks, prompt engineering with GPT-3 was explored to improve model performance (Brown et al., 2020). Van der Meer et al. employed GPT-3 for few-shot classification by designing prompts that structure topics, premises, and conclusions (Van der Meer et al., 2022). However, crafting prompts can be labor-intensive and limit generalizability, potentially leading to suboptimal results across argument domains.

To address inconsistency, we aim to utilize **Recurrent GPT (RGPT)**, based on an adaptive boosting framework (Zhang et al., 2024). RGPT recurrently ensembles base learners, reducing sensitivity to prompt variations and enhancing consistency. This technique is particularly beneficial for argument-mining tasks where context can greatly affect classification accuracy.

BERT-based large language models have also seen tremendous success. BERT employs bidirectional training, transformers, and self-attention to understand contextual relationships (Devlin et al., 2018). The Robustly optimized BERT (RoBERTa) approach refined performance by optimizing hyperparameters, removing next sentence prediction, training with larger batches, and dynamically adjusting the masking pattern (Liu et al., 2019).

In sentence pair regression tasks like validity and novelty assessment, context is crucial. However, BERT and RoBERTa typically incur high computational costs by jointly processing premises and conclusions. Nils Reimers and Iryna Gurevych addressed this by developing Sentence-BERT (SBERT) to accelerate pair-finding while maintaining BERT's performance (Reimers and Gurevych, 2019). Team NLP@UIT used this strategy to improve results marginally over a newer version of RoBERTa. The 2021 RoBERTa adopts a three-stage paradigm (pre-training, post-training, fine-tuning), which proved robust across NLP benchmarks (Zhuang et al., 2021).

Though many teams targeted premise-conclusion connections, the topic of an argument should not be overlooked. A person could obtain insights regarding the validity of a conclusion given its topic alone. Abels et al. employed topic modeling to construct sparser knowledge graphs and more relevant paths, thereby enhancing argument-related properties (Abels et al., 2021). However, we plan to incorporate topic-conclusion relations via incorporating **SBERT** topic embeddings to our neural models.

The best-performing teams in the shared task employed multiple approaches. Team CLTeamL used GPT-3 with fine-tuned RoBERTa, achieving the highest macro F1 score (Van der Meer et al., 2022). Team AXiS@EdUni combined WikiData knowledge graphs, SBERT, and BART via a neural network (Saadat-Yazdi et al., 2022). Team ACCEPT leveraged feature-based SVMs with ConceptNet and SBERT.

Building on these methods, we propose two modeling strategies: feature-based models that combine a domain-specific SBERT model and WikiData features through a neural network, and Fine-tuned GPT-based models that leverage the latest LLM advancements.

3 Dataset

The task data is in American English and contains Premise, Conclusion, Topic, and a Validity and Novel label. The topics present in the training set are distinct from those found in both the validation and test sets. On the flip side, the topics included in the validation set all appear within the test set.

Split	Size	Distribution	Topics
train	750	331 /18 /296 /105	22
validation	202	33 /44 /87 /38	8
test	520	110 /96 /184 /130	15

Table 1: Data Overview. The distribution column indicates the class distribution of (nonvalid, non-novel)/(non-valid, novel)/(valid, nonnovel)/(valid, novel) counts

4 Models

We experiment with two distinct sets of prediction models: the Feature-based model and the Fine-tuned GPT-based Model. Additionally, We separate the validity and novelty prediction tasks and perform binary classifications with each argument quality (valid v.s. non-valid, novel v.s. non-novel).

Our innovative approach involves refining Knowledge Fraph features by creating targeted entities via Named-Entity Recognition and the Wikifier annotation service, as well as incorporating a topic-specific SBERT model for natural language inference (NLI). For GPT-based models, we employ techniques such as Recurrent GPT and Instruction GPT, providing better adaptability and

robustness. We'll evaluate each strategy independently to assess their individual contributions to overall performance.

4.1 Feature-based Prediction Models

4.1.1 knowledge graph using Wiki Data

Knowledge graph (KG) entities and relationship paths potentially represent nuanced semantic and logical connections between a premise and conclusion. Thus, We modify the approach of KEViN and extract WikiData entities that are more concise and relevant to the underlying text. For entity extract, we use Wikifier, a web service that takes text as input and returns the correlated Wikipedia concepts (Brank et al., 2017). We feed a premise or conclusion into Wikifier with an aggressive pruning threshold to obtain a list of entities. Additionally, we apply named entity recognition on our text and save the opinionated words and specific mentions as a list of NER entities. The intersection between Wikifier annotations and NER entities becomes our entity list for a given text. We notice the NER entities are conducive to our analysis, particularly with more involved topics. For example, from the conclusion "The settlement of Iraq is not a success", we were able to successfully capture "Iraq" via NER analysis in addition to our Wikifier annotations "2003 invasion of Iraq" and "Settler colonialism".

We then implement the two most significant features from KEViN: **Irrelevancy** and **Average Distance** (Saadat-Yazdi et al., 2022). They are constructed from KG paths between the premise-conclusion entities. Here, Irrelevancy captures how many conclusion entities cannot be connected with premise entities under max depth limit of 4. Average Distance measures how far all pairs of premise and conclusion concepts are located within WikiData.

4.1.2 Cosine Similarity

For each premise and conclusion pair, we obtain the contextual embedding $\vec{x}$, and $\vec{y}$ respectively using a sentence-transformers model. The SBERT model we are using is specifically trained on various natural language inference datasets and built on RoBERTa-large, which fits into the framework of argument mining (USC-ISI, 2020). With the embedding representations, we assess the relevancy between premise and conclusion through cosine similarity given by the equation:

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|\|\vec{y}\|} \quad (1)$$

4.1.3 Predicted Probability

In addition, we concatenate the topic, premise, and conclusion into a single comprehensive string, using the special character "[SEP]" as a delimiter to distinguish the joining points. Using the same model, we extract embeddings and treat each entry as a distinct feature, which are then inputted into a multi-layer perceptron (MLP) network. This process yields the probability of each entry being valid or novel.

4.1.4 Combined Model Construction

With some or all of the aforementioned constructions, we employ a final multi-layer perceptron network to predict validity and novelty independently.

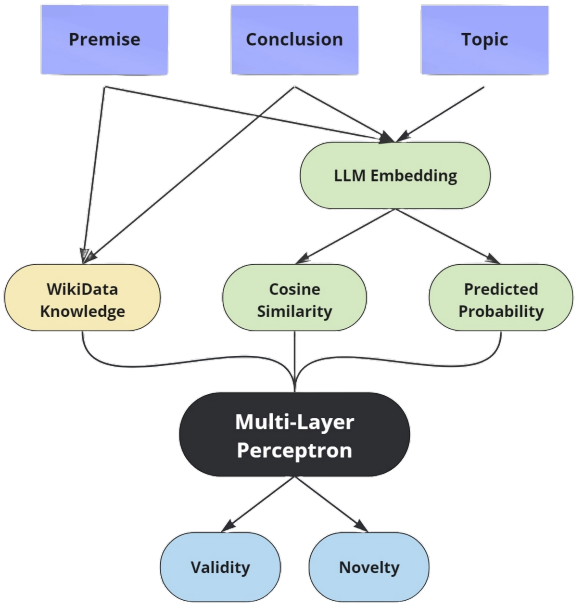


Figure 1: Architecture of Most Complex Model

4.2 Fine-tuned GPT-based Prediction Models

Large language models can be fine-tuned with examples of the instructions, prompts, and responses to make it easier to handle complex tasks. It motivates us to utilize fine-tuned GPT models in an innovative way to improve the predicting performance. We aim to assess the effectiveness of fine-tuned GPT-based models like **Instruction-tuned GPT** in the argument mining task and implement a novel adaptive boosting framework called **Recurrent GPT** for fine-tuning GPT-based models to achieve superior performance.

4.2.1 Instruction-tuned GPT with Few-Shot Learning

Instruction-tuned GPT is a refined version of the traditional GPT architecture that integrates specific instructions and responses. This adaptation allows the model to generate targeted responses that align closely with desired tasks you want them to complete, making it more performant in various applications without complex prompt engineering.

The gpt-3.5-turbo-instruct model is an advanced instruction-tuned model published by OpenAI which is designed for instruction-following tasks. We first try the zero-shot prompts that we are asking the model to complete a task without any examples. Then we apply the few-shot prompts to improve the performance of the model by provide detailed examples.

4.2.2 Recurrent GPT

The original LLM's performance, enhanced through prompt engineering, is constrained, and the full potential of LLMs for classification performance remains largely unexplored. Inspired by the boosting framework, the base learner extends beyond simple models to encompass strong learners such as GPT to achieve optimal performance. Therefore, to overcome the limitations, we implement the **Recurrent GPT (RGPT)** which recurrent ensembles strong base learners through adjusting the distribution of training samples and fine-tuning LLMs. This approach can significantly improve model performance and generalization.

RGPT is an adaptive boosting framework by recurrently fine-tuning and ensembling GPT. It consists of three key steps as follows:

Step 1: Initialization of Base Learner. Train the base model using the raw training sample, and select a general LLM as the initial base learner GPT_0

Step 2: Iteratively fine-tune K base learners, denoted as GPT_k from 1 to K. After each iteration, similar to the adjustment of weights of samples in boosting, the number of samples is adjusted based on the performance of the preceding learner. Specifically, training samples misclassified by the previous learner GPT_{k-1} are given greater weight in the training data, leading to an augmentation of these samples in subsequent iteration. The detailed weight update rule adheres to the formula in this paper (Zhang et al., 2024). Moreover, as shown in Figure 2, the performance and the predicted result of the preceding learner will be incorporated into the input prompt for the fine-tuning of the subsequent learner.

Step 3: Integration of the K fine-tuned learners for inference. Initially, using the original input, the GPT_0 makes the first prediction. Then in a recurrent ensembling way, each subsequent learner GPT_k incorporates the result of the preceding learner GPT_{k-1} into the prompts and generates a new prediction. Finally, the result from the last learner, GPT_K, is considered as the final prediction.

The chain-like structure for fine-tuning in RGPT ensures that the subsequent learners can be improved based on the knowledge learned by the previous learners. It leads to a knowledge accumulation effect and is able to create a more robust GPT at the end of the fine-tuning stage.

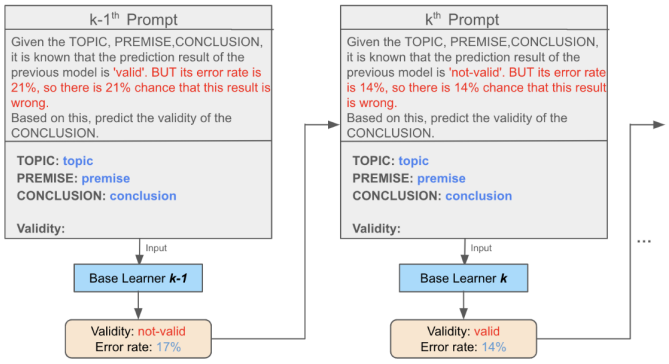


Figure 2: RGPT: Recurrent Ensembling of GPT_k

5 Experiment Setting

5.1 Evaluation Metric

Following the guidelines of ArgMining 2022, we adopt the conventional F_1 score as the metric for assessing the prediction performance. Given the dual nature of our predictions (validity and novelty), the **Macro F1-score** is employed to encapsulate the four metrics combinations and serve as the dominant evaluation metric. The individual F1 scores for validity and novelty are also reported separately.

5.2 Experiment

We trained all models using the 750 training data. Subsequently, we conducted model selection and fine-tuning of hyperparameters with the 202 validation data. Finally, we assessed and compared their performances using the 520 test data. For all the

Feature-based Prediction Models, we apply grid-search in the training process. For GPT-based Prediction Models, we fined-tune and evaluate them using the OpenAI official API. The "gpt-3.5-turbo-instruct" model is used for the instruction-tuned model and "gpt-3.5-turbo" model is used for the RGPT model.

To compare the performance of our methods with other models, we trained the SVM using sklearn to evaluate validity and novelty, creating a competitive baseline. In addition, we use the evaluation results provided by the shared task paper (Heinisch et al., 2022) for a comprehensive comparison with the existing models.

6 Results and Analysis

6.1 Feature-based Results

Approach	F1 Scores		Macro F1
	Validity	Novelty	
Majority Baseline	0.38	0.36	0.130
Embedding only	0.44	0.46	0.221
KG only	0.53	0.36	0.205
KG + Cos Similarity	0.60	0.35	0.235
KG + Embedding	**0.48**	**0.51**	**0.249**
KG + Cos Similarity + Predicted Prob	0.44	0.36	0.180

Table 2: Comparison of F1 Scores for Feature-based Approaches

Overall, features from the WikiData knowledge graph provide model improvements in validity prediction while LLM embedding provide improvement in Novelty prediction. The performances in table 2 reveal a distinct disparity between the tasks of assessing validity and predicting novelty. Our models typically demonstrate better performances in validity tasks compared to novelty tasks. Yet in the best model, novelty reaches a higher F1 score in novelty (0.51) than in validity (0.48) due to a higher embedding dimension. This observation signifies a pronounced trade-off between predicting validity and novelty. The challenge in predicting novelty appears to be inherently linked to the difficulty of creating genuinely novel arguments, as evidenced by the scarcity of novel conclusions within the dataset (around 17% of training data). The class imbalance issue complicates the training of models to effectively recognize and predict novel arguments. On the other hand, the most complex model underperforms when all features are inputted. Given that both cosine similarity and predicted probability are derived from the same embedding, an unimproved outcome is somewhat expected. More critically in retrospect, the predicted probability from the embedding-only model was inherently weak as a feature. Therefore, introducing it as a feature likely introduced additional noise rather than enhancing the model's performance.

6.2 GPT-based Results

Model	F1		Macro F1
	Validity	Novelty	
zero-shot Prompting	0.38	0.30	0.10
Few-Shot Prompting	**0.65**	**0.59**	**0.38**

Table 3: GPT3.5-Turbo-Instruct Results

Table 3 shows the performance of the Instruction GPT with or without few-shot prompting. By providing some examples in the prompt, there is a marked improvement in all metrics when employing Few-Shot Prompting, particularly with the 4-shot configuration in the experiment. This underscores the effectiveness of combining few-shot learning and instruction GPT in enhancing performance.

Model	F1		Macro F1
	Validity	Novelty	
GPT-3.5-Turbo	0.38	0.34	0.1107
SVM baseline	0.53	0.52	0.2657
Instruction GPT	0.65	**0.59**	0.3836
RGPT, k=2	**0.80**	0.49	**0.3843**

Table 4: GPT-based Models Results

From Table 4, it's evident that RGPT with $K = 2$ achieves impressive performance in both the Validity and Novelty tasks. With the highest Validity F1 Score and Macro F1 score, it surpasses other GPT-based models and the SVM baseline. Without the RGPT framework, the original GPT-3.5-Turbo cannot beat the SVM baseline. Figure 4 in the Appendix illustrates a performance increase as the number of base learners increases from 0 to 2. Then the performance remains stable at a very high level. It underscores that the Recurrent framework can enhance the performance of the GPT models in a specific task. The early convergence in performance could be attributed to the small size of the dataset, which may lead to overfitting if the number of learners keeps increasing. Further, we can see the Instruction GPT with few-shot learning has the highest F1 score in the Novelty task. This finding motivates us to propose a combination of Instruction GPT and RGPT to develop a robust

final model. Specifically, we use RGPT for the Validity task and Instruction GPT for the Novelty task.

6.3 Comparison with the state-of-the-art

We present a holistic overview of the model performance in relative to the established baseline and other competitors. As illustrated in Table 5, our feature-based model surpasses the fine-tuned RoBERTa's performance. This shows the effectiveness of the feature-based model in our task. Even more impressively, our RGPT + Instruction GPT model surpassed the top 3 state-of-the-art models on the 2022 shared task's leaderboard in terms of Marco F1, securing the highest ranking. The impressive performance underscores the considerable capability of fine-tuned GPT-based models in our argument-mining task compared to existing advanced models.

Model	F1		Macro F1
	Validity	Novelty	
RGPT + Instruction GPT	0.80	0.59	**0.4784**
CLTeamL-3	0.61	0.75	0.4516
AXiS@EdUni-1	0.70	0.62	0.4327
ACCEPT-1	0.59	0.70	0.4312
Feature-basedKG + Embedding	0.48	0.51	0.2487
RoBERTa Baseline	0.60	0.36	0.2390

Table 5: Comparison with the state-of-the-art

7 Conclusions

In this paper, we demonstrate that features derived from knowledge graph construction and an NLI-focused SBERT model offer distinct predictive signals. Higher-dimensional embeddings significantly enhance the prediction of novelty, whereas knowledge graph and cosine similarity features are more conducive in determining valid arguments. This provides insight for future argument-mining studies on the construction and selection of features.

The more intriguing result lies in the fine-tuned GPT-based model, where combining RGPT and Instruction GPT together reach the highest macro F1 score among existing models in the leaderboard. The RGPT achieved state-of-the-art performance through an adaptive boosting framework, providing a significant improvement in validity prediction. Meanwhile, Instruction GPT excelled at novelty prediction by leveraging few-shot learning. The evaluation results highlight the huge potential of the fine-tuned RGPT framework in tackling other generalized NLP tasks.

Both frameworks warrant further research. In feature-based models, researchers could experiment with various network architectures or revert to classical machine-learning techniques for classification. Additionally, other relevant knowledge graph features, such as entity type or frequency, could help uncover novel sentence relationships. In the fine-tuned GPT-based model, RGPT still has limitations like high computational cost and potential overfitting issues due to its repeated fine-tuning process. Future studies could explore low-cost fine-tuning methods and examine the impact of different dataset sizes on RGPT model performance.

References

Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Focusing knowledge-based graph argument mining via topic modeling. *arXiv preprint arXiv:2102.02086*.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, Ljubljana, Slovenia.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Philipp Heinisch, Anette Frank, Juri Opitz, Moritz Plenz, and Philipp Cimiano. 2022. Overview of the 2022 validity and novelty prediction shared task. In *Proceedings of the 9th Workshop on Argument Mining*, pages 84–94, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

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 preprint arXiv:1907.11692*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Ameer Saadat-Yazdi, Xue Li, Sandrine Chaussón, Vaishak Belle, Björn Ross, Jeff Z. Pan, and Nadin Kökciyan. 2022. Kevin: A knowledge enhanced validity and novelty classifier for arguments. pages 104–110.

Push Singh. 2002. The public acquisition of commonsense knowledge. *In Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.

USC-ISI. 2020. sbert-roberta-large-anli-mnli-snli.

Michiel Van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. 2022. Will it blend? mixing training paradigms prompting for argument quality prediction. 2022:95–103.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. pages 1218–1227.

8 Appendix

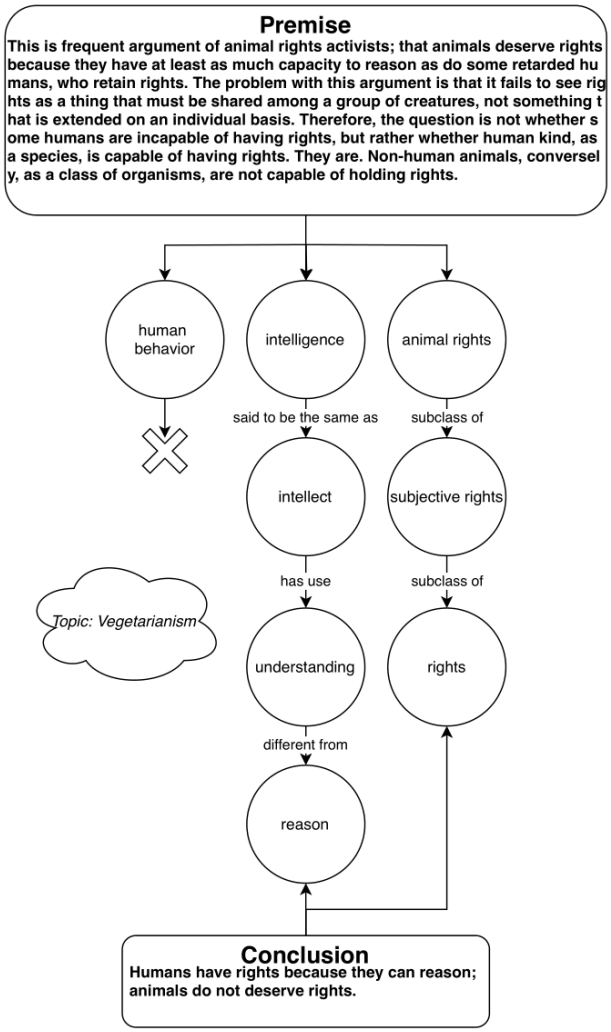
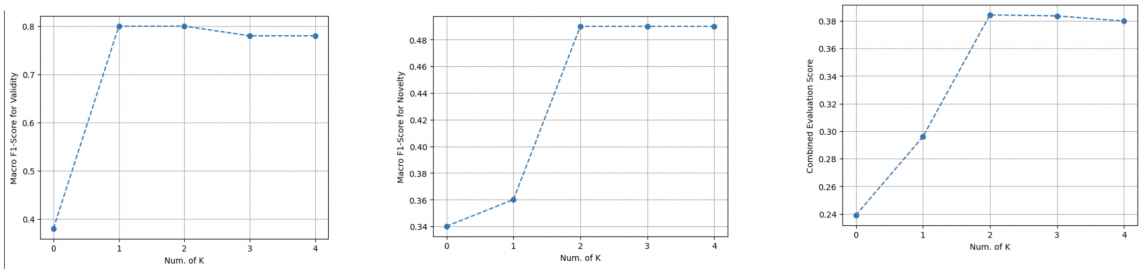


Figure 3: Sample of shortest paths from a premise-conclusion pair.



(a) Validity F1 (b) Novelty F1 (c) Macro F1

Figure 4: Performance of RGPT with increasing number of learners