Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction

Michiel van der Meer
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Myrthe Reuver
Vrije Universiteit Amsterdam
myrthe.reuver@vu.nl

Urja Khurana
Vrije Universiteit Amsterdam
u.khurana@vu.nl

Lea Krause
Vrije Universiteit Amsterdam
l.krause@vu.nl

Selene Báez Santamaría
Vrije Universiteit Amsterdam
s.baezsantamaria@vu.nl

Abstract

This paper describes our contributions to the Shared Task of the 9th Workshop on Argument Mining (2022). Our approach uses Large Language Models for the task of Argument Quality Prediction. We perform prompt engineering using GPT-3, and also investigate the training paradigms multi-task learning, contrastive learning, and intermediate-task training. We find that a mixed prediction setup outperforms single models. Prompting GPT-3 works best for predicting argument validity, and argument novelty is best estimated by a model trained using all three training paradigms.

1 Introduction

As debates are moving increasingly online, automatically processing and moderating arguments becomes essential to further fruitful discussions. The research field of automatic extraction, analysis, and relation detection of argument units is called Argument Mining (AM, Lawrence and Reed, 2020).

The shared task of the 9th Workshop on Argument Mining (2022) focuses on argument quality (Wachsmuth et al., 2017). Argument quality can be broken down into multiple dimensions, each with its own purpose, or be extended to *deliberative quality* (Vecchi et al., 2021). The shared task includes two aspects of the *logical* argument quality dimension: *validity* and *novelty*. Given a premise and a conclusion, a valid relationship indicates that sound logical inferences link the premise and conclusion. A novel relationship indicates that new information was introduced in the conclusion that was not present in the premise. Prediction of an argument's validity and novelty can be either through binary classification (Task A) or by explicit comparison between two arguments (Task B). We focus on Task A.

A system that is able to estimate validity and novelty could be a building block in AM for online deliberation. For instance, in assisting humans to detect arguments in online deliberative discussions (van der Meer et al., 2022; Falk et al., 2021) or presenting diverse viewpoints to users in a news recommendation system (Reuver et al., 2021a).

We address the task of validity and novelty prediction through a variety of approaches ranging from prompting, contrastive learning, intermediate task training, and multi-task learning. Our best-performing approach is a mix of a GPT-3 model (through prompting) and a contrastively trained multi-task model that uses NLI as an intermediate training task. This approach achieves a combined Validity and Novelty F1-score of 0.45.

2 Related Work: Paradigms & Prompting

Given the two related argumentation tasks (novelty and validity), a Multi-Task Learning (MTL) setup (Crawshaw, 2020) is a natural approach. Multi-task models use training signals across several tasks, and have been applied before in argument-related work with Large Language Models (LLMs) (Lauscher et al., 2020; Cheng et al., 2020; Tran and Litman, 2021). We use shared encoders followed by task-specific classification heads. The training of these encoders was influenced by the following two lines of work.

First, intermediate task training (Pruksachatkun et al., 2020; Weller et al., 2022) fine-tunes a pre-trained LLM on an auxiliary task before moving on to the final task. This can aid classification performance, also in AM (Shnarch et al., 2022).

Second, contrastive learning is shown to be

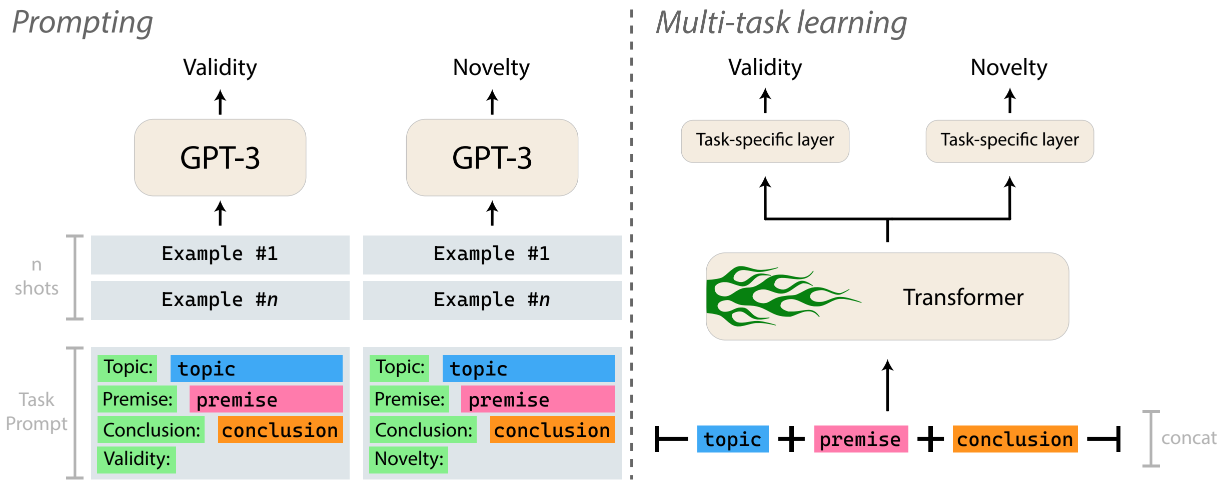


Figure 1: The two argument quality prediction setups used in our submissions. At inference time, predictions from different approaches may be mixed.

a promising approach (Alshomary et al., 2021; Phan et al., 2021) in a previous AM shared task (Friedman-Melamed et al., 2021). Contrastive learning is used to improve embeddings by forcing similar data points to be closer in space and dissimilar data points to be further away. Such an approach may cause the encoder to learn dataset-specific features that help in downstream task performance.

In addition to MTL, we look at prompt engineering for LLMs, which has shown remarkable progress in a large variety of tasks in combination with (Brown et al., 2020) or without few-shot learning (Sorensen et al., 2022). For this task we draw inspiration from ProP (Alivanistos et al., 2022), an approach that ranked first in the “Knowledge Base Construction from Pre-trained Language Models” challenge at ISWC 2022.[1] ProP reports the highest performance with (1) larger LLMs, (2) shorter prompts, (3) diverse and complete examples in the prompt, (4) task-specific prompts.

3 Data and Training Paradigms

3.1 Data

The task data is in American English and consists of Premise, Conclusion, Topic, and a Novel and Validity label. As highlighted in Table 1, arguments that are both non-valid and novel are underrepresented in the data. We use the original training and validation distribution as provided and do not use any over- or undersampling strategies. Instead, we opt to resolve the data imbalance by adopting different training paradigms (see Section 3.2).

Split	Size	Distribution	Topics	Topic Overlap	
				w. train	w. dev
train	750	331/18/296/105	22	–	0
dev	202	33/44/87/38	8	0	–
test	520	110/96/184/130	15	0	8

Table 1: Shared task data overview. **Distribution** indicates the class distribution of {non-valid, non-novel}/{non-valid, novel}/{valid, non-novel}/{valid,novel} counts. The red count indicates a severe data imbalance in the training set.

The content included in the dataset concerns common controversial issues popular on debate portals (Gretz et al., 2020), with topics varying from “TV Viewing is Harmful to Children” to "Turkey EU Membership”.

The training data also contains classes labelled “defeasibly” valid and “somewhat” novel, which are not in the development or test set. We map these to negative labels (i.e. not novel or not valid) to refrain from discarding data. However, we do not measure the effect of this decision on performance.

3.2 Training Paradigms

In our submissions, we mix different training paradigms to obtain our final approach. A schematic overview is given in Figure 1. Below, we outline each of the paradigms individually.

Multi-task Learning Since both validity and novelty are related, a shared encoder is used to process the text input into an embedding, which is fed to task-specific layers. We do not use any parameter freezing, allowing gradients from either task to pass through the entire encoder. During

[1] LM-KBC, https://lm-kbc.github.io/

training, a single task is sampled uniformly at random, and a batch is sampled containing instances for that task.

Intermediate task training In our case, we use two related tasks for intermediate task training: Natural Language Inference (NLI) and argument relation prediction. For NLI, we use a released RoBERTa model (Liu et al., 2019) trained on the MNLI corpus (Williams et al., 2018), predicting whether two sentences show logical entailment. This is related because making sound logical inferences plays a role in validity. The released argument relation RoBERTa model (Ruiz-Dolz et al., 2021) was trained on the relationship (inference, contradiction, or unrelated) between two sentences in a debate (Visser et al., 2020). This is related to novelty and validity. For instance, unrelated arguments may be novel but not valid, and vice versa.

Contrastive Learning We use SimCSE's (Gao et al., 2021) supervised setting to further fine-tune the previously mentioned RoBERTa MNLI model in a contrastive manner. To train the model we take triples of premises and conclusions in the form of premise, conclusion with a positive novelty rating, and conclusion with a negative novelty rating.

4 Approach

4.1 Submitted Approaches

Approach 1: GPT-3 Prompting In our prompt-engineering approach, we use OpenAI's GPT-3[2] (Brown et al., 2020) for few-shot classification of novelty and validity labels. We construct a prompt by concatenating the topic, premise, and conclusion in a structured format, and request either a validity or novelty label in separate prompts. In addition, we show four static examples before asking for a label from the model, selected from short, difficult examples (i.e. those with the lowest annotation agreement) in the training dataset.

Approach 2: NLI as Intermediate-task, Contrastive learning and Multi-Task Learning This model consists of a shared encoder with task-specific classification heads. We initialize the shared encoder using a pretrained RoBERTa model on the MNLI corpus. We then perform contrastive learning with a triplet loss. Afterward, the model is fine-tuned using MTL on the shared task training data. During training, we switch uniformly at random during training between the novelty and validity tasks.

Approach 3: Mixing Approach 1 (GPT-3) & Approach 2 (NLI+contrastive+MTL) Our Mixed Approach uses Approach 1 (prompt engineering) for validity labels, and Approach 2 (fine-tuned model) for novelty labels.

Approach 4: ArgRel as Intermediate-task and Multi-Task Learning This model uses intermediate-task training on the argument relation prediction task followed by Multi-Task Learning in the same set-up as in Approach 1, but without contrastive learning.

Approach 5: Mixing Approach 1 (GPT-3) & Approach 4 (ArgRel+MTL) This approach uses Approach 1 (prompt engineering) for validity and Approach 4 (ArgRel+MTL) for novelty labels.

4.2 Non-submitted Approaches

Baseline: SVM Support Vector Machines (SVMs) are strong baselines for argument mining tasks with relatively small multi-topic datasets (Reuver et al., 2021b). We train an SVM separately for validity and novelty as a competitive baseline.

4.3 Implementation details

We use Python3 and the HuggingFace `transformers` (Wolf et al., 2020) framework for training our models. The SVM baseline instead uses sklearn (Pedregosa et al., 2011). Our code is publicly available.[3] All models trained use RoBERTa (large) (Liu et al., 2019) as the base model, and the intermediate task trained models are obtained directly from the HuggingFace Hub.[4] We provide hyperparameters for fine-tuned trained models in Appendix A.

Model selection was done based on the combined (validity and novelty) F1 performance on the development set. All experiments were run for 10 epochs, after which the best-performing checkpoint was selected for use in creating predictions on the test set. The training was performed on machines including either two GTX2080 Ti GPUs, or four GTX3090 GPUs.

[2] https://beta.openai.com/playground

[3] https://github.com/m0re4u/argmining2022

[4] https://huggingface.co/

Model	F1		
	Validity	Novelty	Combined
SVM (TF-IDF + stemming)	0.60	0.08	0.21
GPT-3 (CLTeamL-1)	0.75	0.46	0.35
NLI+contrastive+MTL (CLTeamL-2)	0.65	0.62	0.39
GPT-3 & NLI+contrastive+MTL (CLTeamL-3)*	**0.75**	**0.62**	**0.45**
ArgRel+MTL (CLTeamL-4)	0.57	0.59	0.33
GPT-3 & ArgRel+MTL (CLTeamL-5)	0.75	0.59	0.43

Table 2: Test set performance. CLTeamL-*n* indicates an official submission with *n* corresponding to the Approach number also in Section 4.1. Bold scores indicate the best-performing approach in the shared task. "Combined" indicates the Shared Task organizer's scoring metric for both tasks.

5 Experiments and Results

We compare our approaches' performance on the test set with the shared task's metric (Combined F1 of Validity and Novelty). Additionally, we analyze our approaches' errors and their connection to labels, annotator confidence, and topic.

5.1 Test set performance

See Table 2 for performance on the test set. We also present a not-submitted SVM as a baseline.

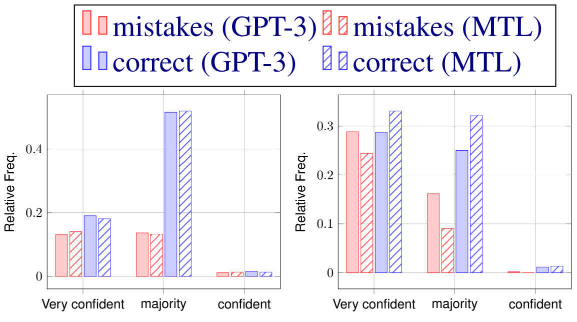
5.2 Error Analysis

We perform additional error analysis on three approaches (Approach 1, 2, and 3). We analyze errors in terms of (1) label-specific performance, (2) annotator confidence, and (3) topics. Additional results are in Appendix B.

Per-label performance We observe complementary strengths for the GPT-3 model and our MTL approach in Tables 3. The MTL model is remarkably stronger than GPT-3 at identifying *novel* arguments, even when considering this is a low-frequency class. We see a similar trend in terms of misclassifications (Table 4), as the MTL model has a 40% lower error rate for the novelty label.

Model	F1 Validity		F1 Novelty	
	valid	non-valid	novel	non-novel
GPT-3	0.78	0.62	0.28	0.67
MTL	0.80	0.50	0.48	0.75

Table 3: Per-label performance on the test set.



(a) Validity

(b) Novelty

Figure 2: Relative accuracy rates divided over label confidence scores.

		Predicted	
		-	+
True	-	237	57
	+	184	42

(a) GPT-3

		Predicted	
		-	+
True	-	265	29
	+	145	81

(b) MTL

Table 4: Confusion matrices for the novelty labels.

Annotator confidence See Figure 2 for the relationship between annotator confidence and classification error. Surprisingly, examples labeled as very confident (easy for human annotators) are not consistently correctly classified by any approach. For novelty, GPT-3 gets about half of these examples wrong.

Topics The 3 topics with the highest error rates differ between approaches and tasks. For validity, GPT-3 struggles with "Was the Iraq War Worth it?" (44.8%), while MTL with "Vegetarianism" (40%). For novelty, GPT-3 also struggles with "Vegetarianism" (60%), and MTL with "Withdrawing from Iraq" (44.7%) and "Vegetarianism" (44%).

6 Conclusion

We highlight two main conclusions.

(1) **Different models have different strengths relating to the two tasks**. A prompting approach with a generative model worked best for validity, while contrastive supervised learning worked best for novelty. The two tasks are related enough to be able to effectively use one multi-task learning model, but merging predictions from multiple heterogeneous models leads to the best score.

(2) **Specific intermediate-tasks before fine-tuning work well for low-resource argument mining tasks**. NLI seems clearly related to validity prediction. For the novelty tasks, other tasks related to argument similarity (Reimers et al., 2019) might be equally informative.

7 Access and Responsible Research

A core consideration in NLP research when sharing results is the accessibility and reproducibility of the solution. While our code is openly available, the approaches including GPT-3 require access to commercially trained models. We used free trial OpenAI accounts (allowing $18 of free GPT-3 credit), but larger datasets and additional tasks can quickly make this approach infeasible. We also considered the freely accessible LLM BLOOM[5]. BLOOM does not require payment, but does require more GPU memory than what was available to us – making it inaccessible.

Ultimately, GPT-3 and related LLMs have several biases and risks of use, including the generation of false information (Tamkin et al., 2021) and the fact that their training on internet language leads to a very limited set of language, ideas, and perspectives represented (Bender et al., 2021), with even racist, sexist, and hateful views (Gehman et al., 2020). This is especially important to mention, as the task description mentions a future use case of generating new arguments.

[5] `https://huggingface.co/bigscience/bloom`

Acknowledgements

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) through the *Hybrid Intelligence Centre* via the Zwaartekracht grant (024.004.022), the *Rethinking News Algorithms* project via the Open Competition Digitalization Humanities & Social Science grant (406.D1.19.073), and the *Spinoza* grant (SPI 63-260) awarded to Piek Vossen. We would also like to thank the reviewers for their excellent feedback that enhanced this paper. All remaining errors are our own.

References

Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as probing: Using language models for knowledge base construction.

Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. 2021. Key point analysis via contrastive learning and extractive argument summarization. In *Proceedings of the 8th Workshop on Argument Mining*, pages 184–189.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141.

Roni Friedman-Melamed, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Conference on Empirical Methods in Natural Language Processing*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

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.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Lawrence and Chris Reed. 2020. 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*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hoang Phan, Long Nguyen, and Khanh Doan. 2021. Matching the statements: A simple and accurate model for key point analysis. In *Proceedings of the 8th Workshop on Argument Mining*, pages 165–174.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter van Atteveldt. 2021a. Are we human, or are we users? the role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 47–59.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021b. Is stance detection topic-independent and cross-topic generalizable?-a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56.

Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. Cluster & tune: Boost cold start performance in text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7639–7653.

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. Hyena: A hybrid method for extracting arguments from opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pages 1–15, Amsterdam, the Netherlands. IOS Press.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Orion Weller, Kevin Seppi, and Matt Gardner. 2022. When to use multi-task learning vs intermediate finetuning for pre-trained encoder transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–282.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

A Hyperparameters

GPT-3 Prompt We used the model `text-davinci-002` with a temperature of 0 and no penalties on frequency and presence. We experimented with various prompt designs (e.g. dynamic or longer examples, more/fewer examples, joint prompting of novelty and validity) but manual inspection showed the best results for the present setup described in the paper (i.e. separate prompts, static prompt style).

Transformers We report the hyperparameters for each approach in Table 5 that differ from the default. In all Transformer models, we used the AdamW optimizer (Loshchilov and Hutter, 2018).

Model	LR	epochs	g.acc.
CLTeamL-2	1e-05	9	1
CLTeamL-3 (novelty)	1e-05	9	1
CLTeamL-4	5e-06	6	4
CLTeamL-5 (novelty)	5e-06	6	4

Table 5: Hyperparameters for our approaches that involve gradient-based learning.

SVM The best performing model on the validation set is one with a C parameter of 0.09 for validity and 4.7 for novelty. The text representation concatenates the two texts, in a TF-IDF and stemmed (with the SnowBall stemmer as implemented in NLTK) representation.

B Additional results

For every analysis, we show the results for approaches *CLTeamL-1* and *CLTeamL-2*, which can be combined into *CLTeamL-3* by merging their results (take validity and novelty, respectively for *1* and *2*).

B.1 Per-label Performance

See Tables 6 and 7.

B.2 Label confusion

See Tables 4 and 8.

B.3 Seed Variance

While the results for the task were obtained using a single model, we investigate training stability over multiple seeds. We show the results and variance from five different seeds for our best-performing MTL model. The results can be seen in Figure 3.

	Prec.	Rec.	F1	Support
non-valid	0.583	0.670	0.623	179
valid	0.812	0.748	0.779	341
non-novel	0.816	0.570	0.671	421
novel	0.199	0.455	0.277	99

Table 6: Performance statistics for approach *CLTeamL-1*.

	Prec.	Rec.	F1	Support
non-valid	0.364	0.806	0.502	93
valid	0.943	0.693	0.799	427
non-novel	0.901	0.646	0.753	410
novel	0.358	0.736	0.482	110

Table 7: Performance statistics for approach *CLTeamL-2*.

Training is relatively stable, but individual models may have small performance differences on the test set.

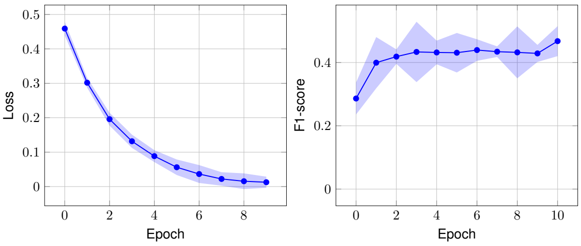


Figure 3: Training loss and combined F1 score for multiple training runs of *CLTeamL-2* with different seeds.

B.4 Topics

The three most error-prone topics were different for approaches. Notable is that “Vegetarianism" is an error-prone topic across tasks and approaches.

GPT-3 - Validity “Was the Iraq War Worth it?" (unseen) with 44.8% errors, “Year Round School" (unseen), 39.7% errors, and “Withdrawing from

		Predicted	
		-	+
True	-	120	86
	+	59	255

(a) GPT-3

		Predicted	
		-	+
True	-	75	131
	+	18	296

(b) MTL

Table 8: Confusion matrices for the validity labels.

Iraq" (unseen), 38.1% errors.

GPT-3 - Novelty “Yucca Mountain nuclear waste" (62.5% error rate), “Vegetarianism" (60% error rate), “Wiretapping in the U.S. (59.2% error rate).

MTL - Validity “Zero Tolerance Law" (42.1%), “Vegetarianism" (40% error rate) and “Yucca Mountain nuclear waste" (37.5% error rate).

MTL - Novelty “Withdrawing from Iraq" (44.7% error rate), “Vegetarianism" (44% error rate), “Wiretapping in the United States" (44% error rate)

Topics not in dev, only in test “Video games’, “Zero tolerance law’, “Was the War in Iraq worth it?’, “Withdrawing from Iraq’, “Year-round school’, “Veal’, “Water privatization’.