Do We Need Large Models for Argument Classification? Revisiting the Role of Model Compression

Filip Gampel[1], Rafał Olszowski[1], Marcin Pietroń[2],

[1]Faculty of Humanities, AGH University of Krakow,
[2]Faculty of Computer Science, Electronics, and Telecommunications, AGH University of Krakow
Correspondence: rolszowski@agh.edu.pl

Abstract

Large language models have improved argument mining substantially, but the associated computational cost complicates deployment, replication, and systematic comparison. We examine how much compression an open-source large language model can tolerate before argument classification quality degrades. Using gpt-oss-20b as the base model, we study pruning with Wanda and post-training quantization under a zero-shot prompting setup. We evaluate compressed variants on three argument-mining resources, namely UKP, Args.me, and ARIES, and contrast their behavior with general language-model benchmarks. The results show a consistent pattern: moderate pruning preserves most of the original performance on argument classification, whereas activation quantization causes larger and more systematic drops. The findings suggest that argument classification is more compression-tolerant than general-purpose evaluation suites, but only up to a point, and they should not be interpreted as evidence that aggressive compression is universally safe. We therefore position compression as a practical way to reduce model cost for argument analysis, while emphasizing that claims about efficiency gains must distinguish between preserved predictive quality and realized runtime speedups.

1 Introduction

Large language models (LLMs) now deliver strong results on many argument-mining tasks, including stance classification, claim detection, and relation prediction. This progress has come with rapidly increasing model size, memory requirements, and inference cost, which makes evaluation harder to reproduce and limits deployment in resource-constrained settings. For argument mining, this tension is especially relevant because many practical uses require running inference over large collections rather than a handful of carefully selected examples.

Despite the practical importance of efficiency, compression has received far less attention in argument mining than in general-purpose LLM evaluation. Most published work focuses either on better prompting strategies or on stronger base models, while relatively little is known about whether argument classification requires the full capacity of a modern open-source LLM. This gap matters because tasks that look fragile on broad reasoning benchmarks may still be stable for narrower classification problems.

Recent studies of LLM-based argument classification also suggest that performance gains do not come only from scaling parameter counts. Pietroń et al. show that compact or moderately sized models can remain competitive when paired with carefully designed inference procedures and prompting strategies (Pietroń et al., 2024, 2025). That observation motivates our question from a different angle: if argument classification is already less dependent on sheer scale than many other LLM tasks, compression may be especially promising here.

This paper studies that question directly. We start from gpt-oss-20b and apply pruning and post-training quantization, then measure the effect of these interventions on both standard LLM benchmarks and three argument-mining datasets. Our goal is not to claim that compression always improves argument classification, but rather to characterize where performance remains stable and where it begins to degrade.

Our contributions are threefold. First, we provide a focused empirical analysis of compression for argument classification across datasets with different label spaces and discourse structures. Second, we show that moderate sparsification is substantially less harmful on argument classification than on several general benchmarks. Third, we identify the limits of this robustness: higher sparsity and especially activation quantization lead to noticeable deterioration, and gains in accuracy do

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 86–93

not always translate into gains in macro-F1.

2 Related Work

Argument mining has evolved from feature-based pipelines to neural and pretrained language-model approaches (Mochales and Moens, 2011; Lippi and Torroni, 2016; Lawrence and Reed, 2020). Representative datasets cover cross-topic argument identification (Stab et al., 2018), large-scale argument search (Ajjour et al., 2019), and persuasive discussion analysis (Chakrabarty et al., 2019). More recent systems build on pretrained encoders and LLM prompting, which has made zero-shot and few-shot argument classification increasingly competitive (Devlin et al., 2019; Wei et al., 2022; Bar-Haim et al., 2017).

Within argument classification specifically, recent work has explored how smaller open models can be strengthened through prompting, auxiliary reasoning steps, and post-processing rather than by relying exclusively on ever larger base models (Pietroń et al., 2024, 2025). Our paper is complementary to that line of research: instead of improving a fixed model with additional reasoning machinery, we ask how much of the model can be removed or quantized before performance deteriorates.

Model compression has a long history in neural NLP and deep learning more broadly. Pruning removes parameters judged unimportant, while quantization reduces numerical precision to lower memory and computation requirements (Han et al., 2015; Hinton et al., 2015). For LLMs, methods such as Wanda and GPTQ show that large models often contain substantial redundancy, at least on broad benchmark suites (Sun et al., 2024; Frantar et al., 2023).

What remains underexplored is whether the same compression behavior holds for argument mining. Argument classification involves structured semantic distinctions, but it is still a constrained prediction problem with small output spaces. That combination makes it plausible that such tasks require less effective model capacity than open-ended generation or broad reasoning benchmarks.

3 Method

3.1 Compression Setup

Let the base model be parameterized by weights

$$\Theta = \{W_1, \ldots, W_L\}, \tag{1}$$

where each $W_l \in \mathbb{R}^{d_l \times k_l}$ denotes a learned linear operator in layer l. Compression aims to reduce the effective storage and computation associated with Θ while retaining predictive quality.

For pruning, we apply a binary mask $M_l \in \{0,1\}^{d_l \times k_l}$ to each layer,

$$\bar{W}_l = M_l \odot W_l, \tag{2}$$

where $\odot$ denotes element-wise multiplication. In our setup, pruning is unstructured rather than structured: individual weights are set to zero according to Wanda's activation-aware criterion (Sun et al., 2024). This distinction matters because unstructured sparsity may preserve accuracy without automatically yielding wall-clock speedups unless the inference stack supports sparse kernels.

For post-training quantization, weights are mapped to a lower-precision representation. Using a uniform b-bit quantizer, a weight matrix W is approximated as

$$Q_b(W) = \Delta \cdot \mathrm{round}\left(\frac{W}{\Delta}\right), \tag{3}$$

where Δ is a scale determined by the dynamic range of W. The quantized weights are then

$$\bar{W} = Q_b(W). \tag{4}$$

In the experiments below, we consider 8-bit weight quantization and a more aggressive 8-bit weight-plus-activation setting. The round-to-nearest (RTN) approach is used. Each layer is quantized independently, and quantization parameters are determined individually for each channel within a single layer. In case of activations, quantization is performed dynamically (parameters are determined separately for each token).

3.2 Base Model and Pipeline

Figure 1 summarizes the pipeline. We use gpt-oss-20b as the base model for all experiments because it is a popular open-source model and has already shown strong argument-classification performance in prior work (Pietroń et al., 2025). Starting from the original checkpoint, we produce pruned variants at multiple sparsity levels using Wanda. We then evaluate these models either without additional quantization, with 8-bit weight quantization, or with 8-bit quantization for both weights and activations. No task-specific fine-tuning is performed.

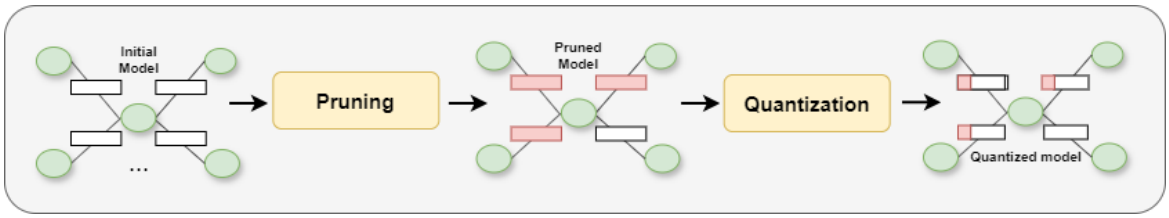

Figure 1: Overview of the experimental pipeline. Starting from gpt-oss-20b, we apply pruning and post-training quantization, evaluate the resulting variants on general and argument-mining benchmarks, and compare how compression affects each task family.

4 Experimental Setup

4.1 Datasets

We evaluate compression on three argument-mining resources chosen to cover distinct formulations of the task.

The UKP corpus (Stab et al., 2018) is a cross-topic argument-mining benchmark built from online comments on controversial issues. We treat it as a three-way classification problem with labels *For*, *Against*, and *No Argument*.

Args.me (Ajjour et al., 2019) contains arguments paired with debated theses from online portals such as Debatewise, Debatepedia, and iDebate. We use it as a binary stance classification task with *For* and *Against* labels.

ARIES (Gemechu et al., 2024) evaluates argumentative relation identification between pairs of argumentative discourse units. We cast it as a three-class classification problem with *Support*, *Attack*, and *No Relation*.

For UKP and Args.me, very long instances were filtered and the evaluation used stratified subsamples of 2,000 instances per subset due computational constraints. Sampling details are summarized in Appendix B.

4.2 Evaluation Protocol

All experiments use fixed task prompts and zero-shot inference. The prompt templates are reported in Appendix B, and the output-parsing summary is reported in Appendix D. For argument-mining datasets, we report accuracy and macro-F1. For the general LLM benchmarks, we report accuracy.

Compression is evaluated at pruning levels of 10%, 20%, 30%, 40%, and 50%. We compare three settings: pruning alone, pruning with 8-bit weight quantization, and pruning with 8-bit weight-plus-activation quantization. The general-benchmark results are included to contextualize how compression affects broad language-model capability relative to argument classification.

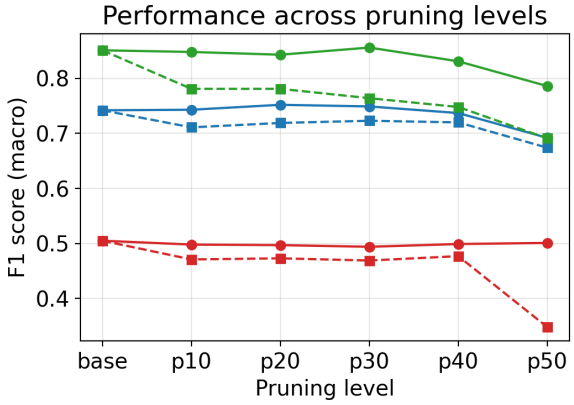

Figure 2: Performance across pruning levels and quantization settings on the argument-mining datasets. Green denotes Args.me, blue UKP, and red ARIES. Solid lines correspond to 8-bit weight quantization; dashed lines additionally quantize activations.

5 Results and Discussion

5.1 General Benchmarks

Table 1 shows that the base model remains broadly stable under moderate compression. Up to 30% pruning with 8-bit weight quantization, most benchmark scores fluctuate only slightly around the baseline. Beyond that point, deterioration becomes more visible, especially on ARC-Easy and ARC-Challenge. The 8-bit weight-plus-activation configuration is clearly less robust even at 40% sparsity.

This pattern is a useful reference for the argument-mining results: once pruning becomes too aggressive, the model's general capability profile degrades.

5.2 Argument Classification

Figure 2 and the detailed results in Appendix A show that argument classification is more compression-tolerant than the general benchmarks, although this robustness is not uniform across

Table 1: Accuracy of pruned gpt-oss-20b variants on general language-model benchmarks.

Model	ARC-Easy	ARC-Challenge	HellaSwag	PIQA	Lambada	Winogrande
base	0.82	0.45	0.58	0.79	0.08	0.67
p10-8w	0.81	0.44	0.57	0.78	0.08	0.69
p20-8w	0.81	0.44	0.60	0.80	0.09	0.67
p30-8w	0.82	0.45	0.58	0.79	0.08	0.65
p40-8w	0.80	0.40	0.58	0.80	0.09	0.68
p50-8w	0.77	0.39	0.60	0.80	0.05	0.69
p40-8w8a	0.73	0.44	0.57	0.73	0.09	0.68

datasets or metrics.

For pruning without additional quantization, UKP is remarkably stable up to 40% sparsity: accuracy stays within 0.005 of the baseline and macro-F1 remains effectively unchanged. Args.me is slightly more sensitive, but the degradation remains small through 30% sparsity and becomes substantial only at 40–50%. ARIES behaves differently. Accuracy rises slightly at higher sparsity, peaking at 0.626 for 50% pruning, but macro-F1 stays essentially flat around 0.50. This means the apparent gain on ARIES should be interpreted cautiously: it likely reflects class-distribution effects rather than a clear improvement in balanced predictive quality.

Adding 8-bit weight quantization preserves the same overall picture. UKP and Args.me remain competitive through moderate sparsity, and ARIES again shows little change in macro-F1 despite small accuracy movements. Weight-only quantization therefore does not materially worsen the compression profile already induced by pruning.

The more aggressive 8-bit weight-plus-activation setting is different. Here, performance drops are systematic across all three datasets. UKP and Args.me lose around four to ten accuracy points depending on sparsity, and ARIES suffers the strongest macro-F1 deterioration at 50% sparsity. The practical conclusion is straightforward: moderate pruning is often acceptable for argument classification, but activation quantization is riskier in this setting.

6 Limitations

The study has several limitations that should temper the conclusions. It covers a single base model family, so the observed robustness may not transfer directly to other architectures. We also evaluate predictive quality rather than direct latency, throughput, or memory savings, which matters because unstructured sparsity does not automatically translate into faster inference on standard hardware. Finally, UKP and Args.me are evaluated on subsamples and the experiments rely on fixed prompts and deterministic answer parsing, which increases uncertainty around small differences.

7 Conclusion

We revisited the role of model compression in argument classification by evaluating pruned and quantized variants of gpt-oss-20b on UKP, Args.me, and ARIES. The central result is that argument classification remains stable under moderate pruning and weight-only quantization even when general LLM benchmarks begin to deteriorate. At the same time, the evidence does not support stronger claims that aggressive compression is universally safe; activation quantization, in particular, introduces substantial risk.

Taken together, the results support a pragmatic conclusion: for argument classification, moderate compression is often a defensible efficiency strategy, but it should be validated with task-specific metrics and runtime measurements.

8 Future works

Several extensions follow naturally from the present study. One direction is task-specific fine-tuning, which would help determine whether compression remains effective once the model is adapted more closely to argument-classification datasets. A second direction is to evaluate multi-prompt inference strategies in which predictions from several prompt formulations are aggregated, as this may improve robustness when individual prompts are sensitive to wording. A third direction is to design prompts that target recurring error sources more directly, particularly contrastive discourse structures, multi-faceted arguments, multiple negations, and cases in which referential alignment is lost.

9 Funding

This research was funded by Narodowe Centrum Nauki (National Science Centre, Republic of Poland), the research grant UMO-2023/49/B/HS5/01379, "Argument Mining: Public Debate Models and Algorithmically-Assisted Argument Extraction" ("Argument mining: Modele debat publicznych i wspomagana algorytmicznie ekstrakcja argumentów"). We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018546.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In C. Benzmüller and H. Stuckenschmidt, editors, *KI 2019: Advances in Artificial Intelligence*, volume 11793 of *Lecture Notes in Computer Science*. Springer, Cham.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *International Conference on Learning Representations*.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marcin Pietroń, Filip Gampel, Jakub Gomułka, Andrzej Tomski, and Rafał Olszowski. 2025. A Comprehensive Study of LLM-Based Argument Classification: from Llama through DeepSeek to GPT-5.2. In preprint.

Marcin Pietroń, Rafał Olszowski, and Jakub Gomułka. 2024. Efficient argument classification with compact language models and chatgpt-4 refinements. In N.T. et al. Nguyen, editor, *Computational Collective Intelligence. ICCCI 2024*, volume 14810 of *Lecture Notes in Computer Science*. Springer, Cham.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 3664–3674. Association for Computational Linguistics.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Table 2: Performance of pruned models without additional quantization.

	UKP		Args.me		ARIES	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
base	0.781	0.742	0.857	0.851	0.605	0.505
p10	0.780	0.744	0.848	0.842	0.601	0.499
p20	0.785	0.751	0.852	0.845	0.597	0.498
p30	0.783	0.747	0.855	0.848	0.591	0.494
p40	0.782	0.744	0.840	0.832	0.610	0.505
p50	0.750	0.694	0.807	0.791	0.626	0.503

Table 3: Performance of pruned models with 8-bit weight quantization.

	UKP		Args.me		ARIES	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
base	0.781	0.742	0.857	0.851	0.605	0.505
p10	0.779	0.743	0.854	0.848	0.601	0.498
p20	0.785	0.752	0.850	0.843	0.596	0.497
p30	0.784	0.749	0.863	0.856	0.589	0.494
p40	0.776	0.737	0.839	0.831	0.608	0.499
p50	0.749	0.692	0.800	0.786	0.624	0.501

A Detailed results

Detailed calculation results are presented in tables 2, 3, 4.

B Prompts

For UKP, `{topic}` is either one of "abortion", "cloning", "death penalty","legalisation of marijuana","stricter gun laws","minimum wage","nuclear energy", "school uniforms", depending on the dataset. In all other cases, data is taken directly and literally from the dataset.

UKP: Is the sentence: "`{sentence}`" an argument for or against `{topic}`, or is it no argument? Return one of the expressions: “For”, “Against” or “No argument”, without any additional commentary.

Argsme: "Is the sentence: "`{sentence}`" an argument for or against `{thesis}`? Return one of the expressions: “For” or “Against”, without any additional commentary."

ARIES: Given the two propositions:
Proposition 1: "`{prop1}`"
Proposition 2: "`{prop2}`"
What is the argumentative relation between Proposition 1 and Proposition 2? Return one of the expressions: "Support", "Attack" or "No Relation", without any additional commentary.

Table 4: Performance of pruned models with 8-bit weight and 8-bit activation quantization.

	UKP		Args.me		ARIES	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
base	0.781	0.742	0.857	0.851	0.605	0.505
p10	0.742	0.711	0.787	0.781	0.585	0.471
p20	0.751	0.719	0.787	0.781	0.588	0.473
p30	0.757	0.723	0.771	0.764	0.581	0.469
p40	0.740	0.710	0.760	0.748	0.589	0.477
p50	0.710	0.654	0.721	0.691	0.611	0.348

C Datasets

For UKP and Args.me, we performed an initial screening of all datasets to crop very long records (>2000 characters in the `argument/sentence` field). Due to computational restraints, for each dataset we used subsamples of 2000 records. These were generated randomly from the full sets, ensuring that the original class imbalance is preserved. Class counts can be found below:

Table 5: UKP class counts

Dataset	**For**	**Against**	**NoArg**	**Total**
abortion	346	418	1236	2000
cloning	465	552	983	2000
death	250	609	1141	2000
gun	471	398	1131	2000
marijuana	474	506	1020	2000
nuclear	339	476	1185	2000
school	362	485	1153	2000
wage	466	446	1088	2000

Table 6: Args.me class counts

Dataset	**For**	**Against**	**Total**
debatepedia	1490	510	2000
debatewise	1179	821	2000
idebate	988	1012	2000

D Answer Parsing

To extract the final classification from the model outputs, we applied regular expressions (Regex) designed to capture the expected label formats while ignoring extraneous text (e.g., "The answer is..."). Below are the regex patterns used for parsing the cleaned model outputs (reasoning, whitespace and punctuation removed, converted to lowercase):

UKP: `r'^\b(for|against|noargument)\b|\b(for|against|noargument)\b$'`
Argsme: `r'^\b(for|against)\b|\b(for|against)\b$'`
ARIES: `r'^\b(norelation|support|attack)\b$'`

E Experimental Setup and Hyperparameters

All local inferences were conducted on the Athena supercomputer at the Academic Computer Centre Cyfronet AGH. The computations were performed on nodes equipped with 8 $\times$ NVIDIA A100 GPUs (40GB VRAM each).

The models were deployed using the vLLM library. We utilized the default vLLM sampling parameters, with the exception of temperature, which was set to 0.6 for all models. The maximum number of new tokens generated was set to 4096. Other hyperparameters such as reasoning effort were also left to the default values.