From Atomic to Complex tasks: Cross-Tasking Improves Zero-Shot Argument Generation and Retrieval

Lynda Tamine Ahmed Rayane Kebir Merveille Dona Codjo
Enzo Pasquies Jose G Moreno
IRIT, Université de Toulouse
Toulouse, France
{first.last}@irit.fr
{first.last}@utoulouse.fr

Abstract

Cross-task generalization mimics human intelligence through the ability to perform tasks by recalling foundational skills acquired previously. In this paper, we argue that argument generation and argument retrieval are complex tasks that could leverage cross-tasking atomic argument mining and argument quality assessment tasks, even if there is no supervision. We empirically demonstrate the rationale behind our claim through the *ArgLLM* framework[1], including a total of 18.9K instruction data using a multi-choice question-answering format, scaling up through multi-tasking and model merging, six natural language argumentation atomic tasks to four complex argument generation and argument retrieval tasks. Our results and analysis, using the backbone Mistral and Llama models, show that cross-tasking in zero-shot settings outperforms base models and is robust to varying strategies, tasks, and model sizes, offering a valuable trade-off between computational cost and task performance.

1 Introduction

Over the past two decades, argument retrieval (AR) and argument generation (AG) have attracted significant attention, largely motivated by practical applications including intelligent personal assistants (e.g., IBM Project Debater (Bar-Haim et al., 2021)) and argument search engines (e.g., args.me (Wachsmuth et al., 2017b)). In their essence, AG and AR follow a process based on three core tasks: (i) extracting *argument units* (e.g., premises, claims) from large-text corpora; (ii) classifying *relations* between argument units (e.g., pro, cons) and between arguments (e.g., support, attack); and (iii) assessing the *quality* of each argument according to specific criteria (e.g., relevance (Bondarenko et al., 2020)). Fostered by the remarkable capabilities of large language models (LLMs), an extensive body of work in natural language argumentation (NLA) has focused on exploring their potential in performing argument unit extraction (Abkenar et al., 2024; Chen et al., 2024a), argument relation extraction (Abkenar et al., 2024; Gorur et al., 2024; Gemechu and Reed, 2024), and argument quality assessment (Gupta et al., 2024; Muti et al., 2024). However, although important, AG and AR have not yet received the research interest they deserve. The literature on leveraging the capabilities of LLMs for AG (Furman et al., 2023; Kao and Yen, 2024; Mouchel et al., 2024; Eskandari Miandoab and Sarathy, 2024) and AR (Dhole et al., 2025; Thakur et al., 2024) highlights their still open challenges: (i) generated or retrieved arguments are long and complex, making human evaluation costly and time-consuming; (ii) model fine-tuning is only slightly beneficial to base model even with increasing model sizes, inducing furthermore, computational and human-annotation costs; (iii) knowledge prompting and in-context learning help AG and AR but performance is critically sensitive to prompt optimization and to the quality of the examples used, thus lowering model robustness.

In this paper, we overcome these limitations with a new perspective. Specifically, we come back to the essence of AG and AR by considering them as *complex tasks* which invoke three foundational *atomic tasks*: Argument STructure identification (AST), ARgument reLation classification (ARL), and Argument Quality Assessment (AQA). Our key idea draws from previous work in multi-step prompting (Wei et al., 2022; Zhou et al., 2023), cross-task and compositional generalization (Chatterjee et al., 2024; Ye, 2024; Hupkes et al., 2020): ideally, if an LLM has already learned foundational skills, it is expected to be able to solve any task whose outcomes are composable from the combination of these skills, a critical ability akin to hu-

[1]Our data and supplementary material are publicly available at https://huggingface.co/collections/mcodjo/argllm-datasets

Findings of the Association for Computational Linguistics: ACL 2026, pages 24864–24882

man intelligence. With this in mind, we formulate the AG and AR complex tasks as a **cross-tasking learning problem**, whose solution leverages the combination of AST, ARL, and AQA atomic tasks' skills. To this end, we employ **multi-tasking** (Raffel et al., 2020; Radford et al., 2019) and **model merging** strategies (Jang et al., 2024; Wortsman et al., 2022) **to combine the foundational skills**. As a first step in investigating this direction, our goal is not to study how AG and AR can be conceptually decomposed into atomic tasks[2]; instead, **our goal is to empirically evaluate the extent to which the atomic tasks' skills help improve AG and AR complex tasks in zero-shot**. To this end, we propose *ArgLLM*, the first argumentation framework that allows scaling from atomic tasks to AG and AR complex tasks. It includes an instruction-tuning dataset, with a total of 18.9K instruction data, 6 cross-tasking strategies, and a set of 12 LLM-based models for AG and AR. To the best of our knowledge, our proposal brings a new perspective to computational argumentation (CA). It enjoys several potential benefits. First, it improves LLM base models without fine-tuning the complex AG and AR tasks, and does not require knowledge-intensive prompt optimization. Second, we expect LLMs to have a strong cross-task generalization ability to diverse AG and AR tasks (e.g., argument essay generation, claim generation) that can be solved by combining foundational argumentation-related skills. Third, it opens up research opportunities: (i) inherently supporting continual learning by adding to backbone LLMs new foundational argumentation-related skills; and (ii) improving their interpretability by supporting their generation with the atomic tasks' intermediate results.

Contributions: (1) We introduce cross-task generalization from atomic argumentation skills to complex tasks in CA; (2) We release *ArgLLM*, a framework for cross-task generalization of AG and AR tasks; (3) We empirically show the potential of our perspective through in-depth experiments and analyses of cross-tasking strategies, model training methods, and cost, which hopefully will open avenues of research in this direction.

[2]This goal is left for future work as mentioned in the conclusion

2 Related work

2.1 Argument generation and argument retrieval.

AG encompasses a wide range of tasks, including counter-argument generation (Lin et al., 2023; Jo et al., 2021; Alshomary et al., 2020) and essay claim generation (Bao et al., 2022; Liu et al., 2021) to cite but a few. AR aims to select a ranked list of candidate arguments that support or attack a controversial topic (Wachsmuth et al., 2017b). Additional tasks, such as retrieval for comparatives and image retrieval for arguments, have been suggested in the Touché evaluation campaign (Bondarenko et al., 2020). However, the emergence of generative information retrieval (Li et al., 2025), closes the gap between AG and AR, allowing for providing generative answers instead of a ranked list of arguments (Dhole et al., 2025; Kiesel et al., 2025). Previous works have shown the potential of LLMs for AG and AR concerning diverse aspects among which the following: (i) LLMs can generate reasonable-quality arguments and counter-arguments but with a huge supervised training and prompting (Lin et al., 2023; Freedman et al., 2024; Mouchel et al., 2024; Eskandari Miandoab and Sarathy, 2024) or high-quality training argumentation training examples (Furman et al., 2023); (ii) LLMs can serve as judges of argument relevance by relying on supervised training using gold human annotations or self-refinement instructions (Lin et al., 2023; Kao and Yen, 2024). Recently, authors in (Stahl et al., 2025; Farzam et al., 2024) explored instruction-tuning of a wide range of argumentative tasks, including argument mining, argument quality assessment, and AG tasks to design domain-specific LLMs in CA. However, our objective here is different; we are interested in whether LLMs have zero-shot cross-task abilities by combining foundational skills (AST, ARL, AQA) to acquire new skills (AG, AR).

2.2 Cross-task generalization of LLMs

It is well-acknowledged that LLMs exhibit strong abilities for multi-tasking a wide range of NLP tasks (Raffel et al., 2020; Radford et al., 2019). However, it has been shown that they struggle in understanding complex instructions or performing complex tasks involving multiple or compositional steps (He et al., 2024; Chen et al., 2024b). One major finding is their limited ability to reuse learned skills to acquire the required skills to solve com-

plex unseen problems, a challenge known as *cross-task generalization* (Chen et al., 2024b; Lin et al., 2022). To tackle this challenge, previous work addressed several approaches improving one design aspect of LLMs: (i) prompting through in-context learning strategies such as chain-of-though (COT) (Wei et al., 2022), Least-to most prompting (Zhou et al., 2023), and compositional prompting (Chen et al., 2024b). However, these methods require huge prompt optimization; (ii) training: the key underlying idea is to endow LLM with the ability to compose vs. decompose complex tasks in less complex ones by leveraging compositional fine-tuning (Zhang et al., 2024, 2021; Bursztyn et al., 2022) or retrieval-augmented training data (Lin et al., 2022). (iii) orchestration of multiple LLMs: this strategy involves several LLMs bringing each knowledge or task skills, leveraged to build a multi-skill model using either multi-agent frameworks (Khot et al., 2021; Du et al., 2024), or model merging, where a single multi-task model without access to the training data, is built upon the parameters of multiple independent fine-tuned models with different skills (Jang et al., 2024; Wortsman et al., 2022).
Drawing insights from this literature, **we initiate cross-task generalization in CA** by investigating whether zero-shot transfer to AG and AR can be efficiently achieved by multitasking or merging the foundational atomic tasks of AST, ARL, and AQA.

3 Problem definition

3.1 Definitions and notations

Atomic task. An atomic task τ, defined for an input space $\mathcal{X}$ to an output space $\mathcal{Y}$ with a parametric function $f^{\tau}(x;\theta^{\tau}) : \mathcal{X} \rightarrow \mathcal{Y}^{\tau}$, requires a relatively low cognitive load to output $y_i \in \mathcal{Y}^{\tau}$ from a given input $x \in \mathcal{X}$. We consider atomic task-specific loss functions as $\mathcal{L}^{\tau}(\cdot,\cdot) : \mathcal{Y}^{\tau} \times \mathcal{Y}^{\tau} \rightarrow R^{+}$.

Complex task. A complex task $\mathcal{T}$ requires a high cognitive load to output $y \in \mathcal{Y}^{\mathcal{T}}$ from a given input $x \in \mathcal{X}$. In this paper, we consider a task as complex when it involves multiple atomic tasks.

Cross-task generalization. Given a set of n tasks $T = \{\tau_1 \ldots \tau_n\}$ with each task τ_i a training dataset $\mathcal{D}_i$ containg a set of n_i examples $\{(x_k^i, y_k^i)_{k=1}^{n_i}\}$, the objective of cross-task generalization is to leverage training data $\mathcal{D} = \bigcup_{i=1}^{n} \mathcal{D}_i$ and transfer knowledge to facilitate learning a seen task ($\tau_i \in T$) on unseen examples or learning an unseen task ($\tau \notin T$).

3.2 Problem formalization.

In this paper, we hypothesize that AG and AR are complex tasks that could leverage knowledge sharing and transfer from the foundational atomic tasks AST, ARL, and AQA tasks. To investigate the rationale behind our hypothesis, we model AR and AG as a cross-task learning problem and use their performance in zero-shot as a proxy to empirically evaluate the effect of cross-tasking the atomic tasks AST, AQA, and ARL.

Let us consider a language model $\mathcal{M}_{\theta}$ that generates from a vocabulary V an output answer $\mathcal{A} = a_1..a_m$ where $a_i \in V$, is an argument statement (or a list of argument statements), to answer query input q (e.g, claim, essay). $\mathcal{M}_{\theta}$ continuously generates, at inference, the tokens of answer statements a_i as follows:

$$a_i = \mathcal{M}_{\theta}(a_i|\theta^{\mathcal{T}}, a_{<i}, q, \theta) \tag{1}$$

where $a_{<i}$ are the previously generated tokens.

We consider $T = \{\tau_{AST}, \tau_{ARL}, \tau_{AQA}\}$, as the three atomic CA tasks defined respectively by the parametric functions $f^{ast}(x;\theta^{ast}), f^{arl}(x;\theta^{arl}), f^{aqa}(x;\theta^{aqa})$ used to train $\mathcal{M}_{\theta}$ based on two strategies of cross-tasking methodologies:

Multi-tasking (MT): Multi-tasking (Raffel et al., 2020; Radford et al., 2019) is a one-stage process which consists of training the language model backbone $\mathcal{M}_{\theta^0}$ on all the task-specific training examples $(x, y) \in \mathcal{D}$ with $\mathcal{D} = \bigcup_{\tau \in T} \mathcal{D}_{\tau}$ using the loss objective:

$$\mathcal{L}(\theta^{mt}) = \min_{\theta^{sh}, \theta^{\tau}} \sum_{(x,y) \in \mathcal{D}} \alpha_{\tau} \mathcal{L}(f^{\tau}(x;\theta^{sh}, \theta^{\tau}), y) \tag{2}$$

where α_{τ} is a task-specific parameter, and θ^{sh} are parameters shared between the atomic tasks AST, ARL, and AQA.

Model merging (MG): Model merging (Jang et al., 2024; Wortsman et al., 2022) is a two-stage process. The first stage consists of fine-tuning independently the language model backbone $\mathcal{M}_{\theta^0}$ on each task-specific' training examples $(x^i, y^i) \in \mathcal{D}_i$ leading to the models $\mathcal{M}_{\theta^i}, \tau_i \in T$. The second stage involves merging the fine-tuned models using a parameter-wise merging strategy. Formally, the goal of model merging is to find the optimal coefficients $\lambda_i^*, i = 1..n$ so that the merged model $\mathcal{M}_{\theta^{mg}}$ can transfer the capabilities of $\tau_i \in T$. θ^{mg} is computed using a core merging function f as:

$$\theta^{mg} = f(\lambda_i^*, \hat{\tau}_i)_{i=1}^{n} \tag{3}$$

where $\hat{\tau}_i$ is the parametric task vector of τ^i computed by using θ^i and θ^0.

4 The *ArgLLM* Framework

In the following, we detail the *ArgLLM* framework designed to address our problem (§ 3).

4.1 Task categories, tasks and datasets

Atomic tasks (τ). We consider a set of argument mining and argument quality assessment atomic tasks ($n = 6$) that fall into 3 categories:

- *Argument Structure identification (AST)*: includes a set of argument mining tasks whose objective is the identification of argumentative structure in natural language text. We consider the following tasks: (1) argument unit segmentation (US) in persuasive essays using the dataset AAE-v2 Essays (Stab and Gurevych, 2014); the input is a text segment and the output a label (Premise, Claim, Major claim); and (2) claim evidence detection (CD) for fact checking using the dataset CheckThat!2022 (Nakov et al., 2022); the input is a claim and the output is a candidate evidence.

- *Argument Relation (ARL)*: includes a range of argument mining tasks whose objective is to identify the relation between argumentative units. We address the following tasks: (3) claim stance classification (SC) using the IBM Stance dataset (Bar-Haim et al., 2017); the input is a pair (Topic, Claim), the output indicates whether the claim supports (Favor) or contests (Against) the topic; and (4) support/attack relation classification (SA) in persuasive essays using the dataset (Stab and Gurevych, 2014); the input is a pair (Premise, Claim) and the output is the type of relation (Support vs. Attack).

- *Argument Quality Assessment (AQA)*: includes a range of tasks whose objective is to evaluate the strength of an argument based on a set of quality-related criteria. We address tasks based on two different criteria: (5) the argument reuse (AU) criterion where the quality of an input argument's conclusion is determined based on its reuse as a premise, as proposed in the Webis-ArgRank-17 corpus (Wachsmuth et al., 2017c); and (6) the cogency criterion based on the objective acceptability (OA) of an input argument as defined in the Dagstuhl-15512 corpus (Wachsmuth et al., 2017a).

Complex Tasks ($\mathcal{T}$). We explore the following AG and AR tasks:

- *Argument Generation (AG)*: The objective of an AG task is to generate from plain texts argumentative text segments. In this work, we consider: (1) argument essay generation using the AEG dataset (Bao et al., 2022), whose objective is to generate a persuasive argumentative essay from a writing automatically; and (2) claim generation using the iDebate corpus (Wang and Ling, 2016), whose objective is to generate a claim given a set of arguments on the same topic. We posit that achieving each of the above AG tasks' objectives involves multiple atomic tasks: argument structure identification (AST) from plain or argumentative texts, identifying the stance of the argument unit toward a given topic (ARL), and appropriately logically incorporating the arguments in a persuasive argumentative text (AQA).

- *Argument Retrieval (AR)*: This task consists of retrieving relevant and acceptable arguments or counter-arguments to a given claim or topic. In particular, we consider the Arguana (Wachsmuth et al., 2018) and Touché2020 (Bondarenko et al., 2020) datasets.

Datasets' statistics are presented in Table 1.

4.2 Building argument instructions

Learning a unique objective loss for $\mathcal{M}_\theta$ ($\mathcal{M}_{\theta^{mt}}, \mathcal{M}_{\theta^{mg}}$) with the same underlying decoding process faces the issue of the heterogeneity of the input-output (i.e., $\{(x_k^i, y_k^i)\}_{i=1}^n$) formats across the atomic tasks. To tackle this issue, we build a unified text-to-text instruction framework as previously done in NLP (Raffel et al., 2020; Radford et al., 2019), enabling standard causal language model training with a cross-entropy loss. To allow fair comparison with the merging cross-tasking methodology, we adopt the same instruction format and loss objective to train $\mathcal{M}_{\theta^i}$. Specifically, we adopt a unified MCQA format, based on a question and a set of answer options with associated symbols (e.g., A, B). This format is effective to endow LLMs with the task knowledge through relevant answer selection while mitigating token representation bias (Robinson and Wingate, 2022). To this end, we approach each atomic task τ_i as either a binary classification task or a multi-label classification task from which we formulate the *question* and answer *options* to fit the MCQA format presented in Appendix A (full examples for each category are presented in Table 5).

- *Question*: For each atomic task $\tau_i \in T$, we transform the task input element x_k^i in the case of

Type	Task Category	Task/Dataset	Train	Val.	Test
Atomic	**A**rgument **ST**ructure (**AST**)	Unit Segmentation (Stab and Gurevych, 2014)	4860	1215	1575
		Claim Detection (Nakov et al., 2022)	3324	307	911
	Argument **Re**Lation (**ARL**)	Stance Classification (Bar-Haim et al., 2017)	831	208	1355
		Support/Attack Classification (Stab and Gurevych, 2014)	2418	605	809
	Argument **Q**uality **A**ssessment (**AQA**)	Argument Reuse (Wachsmuth et al., 2017c)	165	36	36
		Objective Acceptability (Wachsmuth et al., 2017a)	204	52	64
	Total Atomic		11802	2423	4750
Complex	**A**rgument **G**eneration	AEG (Bao et al., 2022)			1003
		iDebate (Wang and Ling, 2016)			2259
	Argument **R**etrieval	Arguana (Wachsmuth et al., 2018)			1406
		Touché2020 (Bondarenko et al., 2020)			49

Table 1: Statistics of the atomic and complex datasets.

ARL and *claim detection*, or the main input element of the input x_k^τ in the case of the AQA task category, into a question (q_k^τ). For the *unit segmentation* atomic task, we use markers as done in Baldini Soares et al. (2019) to identify segment candidates, enabling us to form the input q_k^i based on the sequence labeling of segment tokens as *claim*, *major claim*, or *premise*.

- *Answer options*: For AST and ARL, we consider the candidate answers a_{kj}^i to each question based on the set of unique labels or gold answers in the dataset (e.g., Favor, Against), and provide them as answer options by ensuring that the gold answer a_i^* is among the options. For AQA, we provide the remaining elements of the input as answer options. Subsequently, we form the training examples (x_k^i, y_k^i) by the concatenation of the question q_k and each of the candidate answers a_{kj}^i to form the input X_k^i, while the gold answer a_i^* becomes the supervision y_k^i.

To enable the language models $\mathcal{M}_\theta$ to understand each of the atomic task objectives and establish patterns from training data (x_k^i, y_k^i) to the atomic task τ_i, we append to each training example a template-based natural language instruction I_i of τ_i to let the model $\mathcal{M}_\theta$ learn the probability $p(x, y, \tau_i)$. Detailed templates are presented in Appendix A, Table 6. The resulting *ArgLLM* training dataset $\mathcal{D}$ includes a set of $N = 18,975$ MCQA instructions split into training, validation, and test sets. Finally, we use causal language modeling as the unified loss objective to train the language models $\mathcal{M}_{\theta^{mt}}$ and $\mathcal{M}_{\theta^i}$.

4.3 Cross-tasking methodologies

Multi-task instruction tuning. We explore three strategies to train the model $\mathcal{M}_{\theta^{mt}}$:

- *Mixing* (ArgLLM$_{MX}$): As done in Raffel et al. (2020), ArgLLM$_{MX}$ corresponds to proportionally mixing instructions training data from multiple atomic tasks' categories to guarantee their joint learning in each training batch.

- *Sequential* (ArgLLM$_{SQ}$): Considers the task sequence based on an intuitive order (AST), then on (ARL), and finally on (AQA).

- *Partially-sequential* (ArgLLM$_{PS}$): We explore a hybrid training strategy where the model fully learns a target atomic task and then subsequently learns additional tasks in a multi-task fashion (Raffel et al., 2020).

Model merging. We explore the following core functions f (§ Eq. 2) to train the model $\mathcal{M}_{\theta^{mg}}$:

-*Model Soup (MS)* (Wortsman et al., 2022): Consists of performing a linear combination of input models' weights using a model-wise coefficient. Formally $\theta^{mg} = \sum_{i=1}^{n} \lambda_i \theta^i$, where $\sum_{i=1}^{n} \lambda_i = 1$ and $\forall_i \lambda_i > 0$.

-*Model Ties (Yadav et al., 2023a)*: This model takes into account the interference between different parameters during merging. Formally $\theta^{mg} = \theta^0 + \lambda \hat{\tau}^{mg}$, where $\hat{\tau}^{mg}$ is the task vector using the mean parameter values from the models whose signs are the same as the aggregated elect sign of the top $k\%$ parameters of each task vector θ^i.

- *Model SVD (Stoica et al., 2025)*: This model first aligns, using a Singular Value Decomposition (SVD), LORA task vectors' parameters with full-finetuned ones before model merging. In this paper, we performed the SVD and applied the linear combination of input models as done in the model Soup (Wortsman et al., 2022), i.e., $\theta^{mg} = \sum_{i=1}^{n} \lambda_i \hat{\theta}^i$ where $\hat{\theta}^i$ is computed as the concatenation of the LoRA task' vector parameters' updates.

To build model $\mathcal{M}_{\theta^{mg}}$, *ArgLLM* relies on parameter tuning of θ^i based on the performance of each atomic task and in the selection of λ^*. For the sake of simplicity, we opted for equal λ^* weights when possible; Appendix B provides further details.

5 Experimental Results and Analysis

This section presents the main experiments and results. Appendix B presents the implementation details, and Appendix C details both these experiments and additional ones.

5.1 Experimental Setup

Backbone models. We use two foundation models, Llama-3-8B[3] (AI@Meta, 2024) and Mistral-

[3]meta-llama/Meta-Llama-3.1-8B-Instruct

7B[4] (Jiang et al., 2023). From these backbone models, we derive the following models:

-ArgLLM$_{AST}$ is finetuned on argument structure datasets within the AST task category and excluding all the other tasks.

-ArgLLM$_{ARL}$ is finetuned on argument relation datasets within the ARL task category and excluding all the other tasks.

-ArgLLM$_{AQA}$ is finetuned on argument quality datasets within the AQA task category and excluding all the other tasks.

-ArgLLM$_{MT}$ is finetuned on all the considered datasets within the atomic task categories AST, ARL, and AQA in a multi-task setting.

-ArgLLM$_{MG}$ is obtained by combining the weights of the three atomic task models: ArgLLM$_{AST}$, ArgLLM$_{ARL}$, and ArgLLM$_{AQA}$.

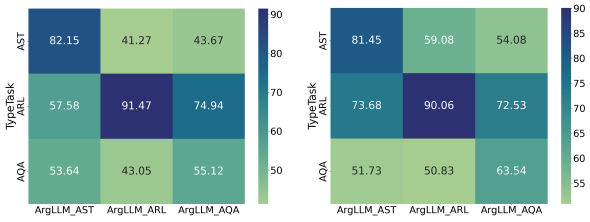
5.2 Effectiveness evaluation on AG and AR

Atomic tasks are worth cross-tasking. First, we aim to understand to what extent the atomic NLA task categories relate. To this end, we conduct a study of "out-of-category" transferability and "task-vector orthogonality" for the two cross-tasking methodologies, namely, multi-tasking and merging, respectively. Each of the atomic-task models ArgLLM$_{AST}$, ArgLLM$_{ARL}$, and ArgLLM$_{AQA}$ was used to build a vector $\hat{\tau}^i$ by using all trainable parameters whose size is the vector dimension. We evaluate the in-task performances as well as the capabilities in transferability. To do so, each model was tested within its own category and in the remaining unseen atomic tasks. Additionally, to measure the task-vector orthogonality, we calculate cosine similarities between each pair of vectors. Figure 1 (a) presents the cosine similarity values between task vectors, while Figure 1 (b) presents the performance results on the atomic tasks. Some expected behavior is to have strong values on the diagonal, paired values by considering each atomic task category, reflecting the "in-category" interdependence of the tasks. While it is the case for most of the tasks for both backbone LLMs, we can see that ArgLLM$_{AQA}$ generalizes to ARL atomic category slightly better than their own category (55.12 vs 74.94 and 63.54 vs 72.53) but the task vectors strongly differ (100 vs 1.24 - 2.11). This observation is consistent with previous findings from Wachsmuth et al. (2024), arguing that LLMs can leverage instruction-following across multiple contexts, but they need to be instructed systematically toward the specific problems to solve. Regarding the "out-of-category" evaluation, we can see that the overall results are significantly lower than the in-category evaluation scores, indicating the out-of-domain knowledge that can be brought by each task category, although the decrease rate is more pronounced with the Llama-3-8B model. Besides, we can observe that models fine-tuned with ARL outperformed on the other tasks. This suggests that this task is usable as a transfer task to other atomic argumentation tasks.

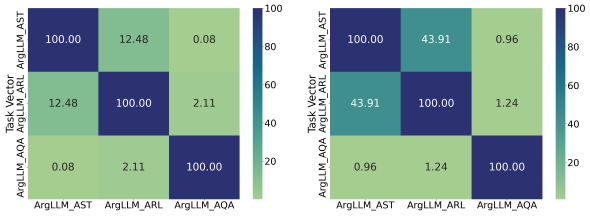
Overall, this experiment emphasizes both task-specific skills and shared information across task categories that could be leveraged through cross-tasking to improve AG and AR.

Cross-tasking atomic tasks improves AG and AR in zero-shot. Our main results for both cross-tasking strategies, using BLEU(-4) and NDCG@10 metrics, are presented in Table 2 (full metrics used for these datasets are presented in Appendix C, Tables 8 and 9). We can see that Llama-3-8B versions of ArgLLM$_{MT}$ and ArgLLM$_{MG}$ outperform several other configurations in the AG complex task. Indeed, in zero-shot configuration for the AG task, Llama-3-8B (4.45 and 2.53) outperforms Mistral-7B (3.73 and 1.75) in both datasets (AEG or iDebate). Moreover, our ArgLLM$_{MT}$ and ArgLLM$_{MG}$ models outperform the zero-shot alternatives regardless of the dataset and models (Llama-3-8B or Mistral-7B), which suggests robustness in the training in atomic tasks. Moreover, some atomic models, such as ArgLLM$_{ARL}$ for Mistral-7B or ArgLLM$_{AQA}$ for Llama-3-8B, outperform the ArgLLM$_{MT}$ and ArgLLM$_{MG}$ models, which suggests that a more elaborate cross-tasking strategy may grasp extra improvement on top of our proposal. Regarding the AR complex task, we can observe from Table 2, that our multi-task strategy outperforms zero-shot capabilities on both datasets (Arguana and Touché2020). More interestingly, the individual atomic tasks ArgLLM$_{AST}$ (35.01), ArgLLM$_{ARL}$ (36.71), and ArgLLM$_{AQA}$ (33.09) outperform the zero-shot Mistral-7B (33.08), but the ArgLLM$_{MT}$ successfully merges the individual efforts up to a 39.26 performance in terms of NDGC@10. This shows that the atomic tasks jointly contribute to outperforming the AR complex task. A similar behavior is observed for the same model (Mistral-7B) on the Touché dataset, but with only ArgLLM$_{AQA}$ slightly outperforming

[4] mistralai/Mistral-7B-Instruct-v0.3



(a) Transferability across atomic tasks.



(b) Task vectors similarity.

Figure 1: Comprehensive analysis of atomic tasks using Mistral-7B (left) and Llama-3-8B (right): Top row (a) shows performance transferability across atomic task datasets, and bottom row (b) depicts the cosine similarities of task vectors.

			Argument Generation		Argument Retrieval	
			AEG	iDebate	Arguana	Touché
			BLEU-4	BLEU	NDCG@10	
ZS		Mistral-7B	3.73	1.75	33.08	39.17
		Llama-3-8B	4.45	2.53	32.14	33.35
Fine-tuned	ArgLLM_{AST}	Mistral-7B	4.60	4.10	35.01	35.37
		Llama-3-8B	4.68	3.50	34.80	31.09
	ArgLLM_{ARL}	Mistral-7B	4.50	**4.30**	36.71	37.60
		Llama-3-8B	4.79	3.40	22.78	34.58
	ArgLLM_{AQA}	Mistral-7B	3.64	3.64	33.09	39.30
		Llama-3-8B	4.38	3.83	32.13	33.40
Cross-Task	ArgLLM_{MT}	Mistral-7B	4.56	2.77	**39.26**	**43.89**
		Llama-3-8B	5.18	3.65	27.94	30.66
	ArgLLM_{MG}	Mistral-7B	5.40	3.80	37.41	38.20
		Llama-3-8B	**5.48**	3.64	33.43	37.27

Table 2: Comparative analysis of *ArgLLM* using BLEU and NDCG metrics in AG and AR complex tasks. Best LLM-based performance is indicated in **bold**.

the zero-shot performances. Results with Llama-3-8B are less successful for ArgLLM_{MT} in AR, however, ArgLLM_{MG} (33.43 and 37.27) shows a different behavior as they outperform the zero-shot baselines of the same model (32.14 and 33.35).

5.3 Ablation study

Table 3 presents the results of our thorough ablation study on the atomic task categories. Our main observations are the following: (1) Mistral-7B version of our ArgLLM_{MT} and ArgLLM_{MG} shows that removing any of the atomic tasks negatively affects the overall performance, except for the iDebate dataset, and, partially, for the Arguana dataset, where removing the ARL atomic task is beneficial in the overall performance. This is consistent on the iDebate dataset with the results in Table 2, where atomic-task models also achieve better results; (2) Llama-3-8B version of our ArgLLM_{MT} is positively affected when removing the AQA atomic task. This specific behaviour of AQA tasks has also been observed in the transferability evaluation of atomic tasks (§ 5.2). It suggests a poor integration of the AQA atomic task into our multi-task strategy for Llama-3-8B that was not observed for ArgLLM_{MT} when using Mistral-7B nor ArgLLM_{MG} (regardless of the LLM backbone); (3) ArgLLM_{MG} is more stable when combining different atomic task as ablating some task may generate slightly better improvements in only 3 out of 8 combinations of LLM and datasets (5.48 vs 5.51, 37.41 vs 39.75, and 37.27 vs 40.13). While this ablation study corroborates the overall benefit of cross-tasking atomic NLA tasks for zero-shot AG and AR, it suggests better tuning their combination within ArgLLM_{MG} and ArgLLM_{MT} to optimize the overall performance. Appendix C (Figure 4) presents a qualitative example for AG using both cross-tasking strategies.

5.4 *ArgLLM* analysis

Impact of cross-tasking strategies. Figure 2 presents comparative results using three Multi-tasking (ArgLLM_{MX}, ArgLLM_{SQ}, and ArgLLM_{PS}) and three Model merging (ArgLLM_{MSoup}, ArgLLM_{SVD}, and ArgLLM_{TIES}) for Mistral-7B and Llama-3-8B. We can see that the multi-task setting, ArgLLM_{MX}, successfully obtains the best performance (4.56, 39.26, and 43.89) over the ArgLLM_{SQ} and ArgLLM_{PS} alternatives for three complex task datasets out of four. Interestingly, by cross-linking the results of the ArgLLM_{SQ} alternative with those of the atomic task models (ArgLLM_{AST}, ArgLLM_{ARL}, ArgLLM_{AQA}) presented in Table 2, we can observe that sequential training is even worse than the models trained on task categories, demonstrating the interest of using a multi-tasking strategy over the atomic tasks to enhance the LLM with AG and AR task understanding. Finally, the ArgLLM_{PS} alternative slightly outperforms the ArgLLM_{SQ} but is still far behind ArgLLM_{MT} for three out of four datasets. A similar pattern is observed with merging models where ArgLLM_{MSoup} outperforms six out of eight merging configurations, with only two exceptions where ArgLLM_{SVD} obtains the best performance for AEG using Mistral-7B and Touché using Llama-3-8B.

Cross Tasking Cost Analysis. Table 4 illustrates the relative performance difference in percentage (raw values are available in Appendix C, Table 11)

Task	Dataset (Metric)	Model	MultiTask (ArgLLM$_{MT}$) -	MultiTask (ArgLLM$_{MT}$) w/o AST	MultiTask (ArgLLM$_{MT}$) w/o ARL	MultiTask (ArgLLM$_{MT}$) w/o AQA	Merging (ArgLLM$_{MG}$) -	Merging (ArgLLM$_{MG}$) w/o AST	Merging (ArgLLM$_{MG}$) w/o ARL	Merging (ArgLLM$_{MG}$) w/o AQA
AG	AEG (BLEU-4)	Mistral-7B	**4.56**	3.63	3.61	3.64	**5.40**	4.81	5.04	5.10
		Llama-3-8B	**5.18**	4.39	4.46	4.45	5.48	**5.51**	5.14	5.04
	iDebate (BLEU)	Mistral-7B	2.77	3.75	**4.01**	3.97	**3.80**	3.69	2.57	3.08
		Llama-3-8B	3.65	3.76	3.85	**3.92**	**3.64**	1.62	1.59	1.62
AR	Arguana (NDCG@10)	Mistral-7B	**39.26**	33.09	33.26	33.67	37.41	33.20	**39.75**	33.67
		Llama-3-8B	27.94	32.21	33.18	**33.50**	**33.43**	33.41	30.53	31.06
	Touché (NDCG@10)	Mistral-7B	**43.89**	37.24	36.28	35.94	**38.20**	37.52	37.17	38.02
		Llama-3-8B	30.66	32.62	**42.60**	36.48	37.27	39.72	**40.13**	37.46

Table 3: Ablation study of ArgLLM$_{MT}$ and ArgLLM$_{MG}$ on atomic tasks.

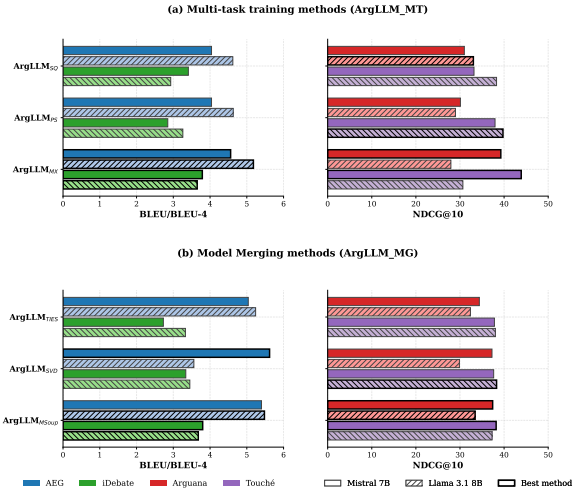


Figure 2: Results of *ArgLLM* versions by using different training and merging strategies.

of *ArgLLM* across three Llama backbone sizes (1B, 3B, 8B), normalized against the 8B vanilla (ZS) baseline. We can see that while increased model size generally improves performance (and computational cost), particularly for generation metrics, we interestingly observe the effectiveness of cross-tasking in smaller models: the 1B *ArgLLM* variants consistently recover or exceed the capabilities of the vanilla 8B baseline. Notably, for the AR task, ArgLLM$_{MT}$-1B (PEFT) achieves similar performance than 8B ZS baseline (-1.8%), despite an 87.5% reduction in total parameters. A similar trend is observed for the 3B variant in the AG task. This suggests that cross-tasking enables smaller models to recover capabilities typically associated with larger backbones.

A second consistent trend that arises from Table 4 is that parameter-efficient fine-tuning (PEFT) matches or outperforms full fine-tuning (FFT) across most settings. For the 1B and 8B scales, PEFT yields substantial optimization margins ($\Delta\%$) over FFT, such as a +41.3% and +26.1% gap in Arguana for MT and MG configurations, respectively. While FFT occasionally shows competitive generation performance at the 8B scale, it is prone to significant regressions at smaller scales. We attribute the robustness of PEFT to the optimization stability of LoRA adapters under limited atomic-task supervision, which mitigates the catastrophic interference and overfitting observed in FFT.

These results support the deployment of our proposal in compute and memory-constrained settings.

Impact of few-shot fine-tuning. We investigate the impact of few-shot fine-tuning using a very small portion (1–10%) of the available training data, on the performance of ArgLLM$_{MG}$ and ArgLLM$_{MT}$. We conduct this analysis on the AEG dataset, which is the only dataset in our setting that provides an official train–test split, allowing fair comparison with baseline models. The full AEG training set consists of 9,276 instances. As shown in Figure 3, with as little as 3% of the AEG training data (278 samples), the Llama-3-8B–based model already surpasses the state-of-the-art baseline (Bao et al., 2022). Similarly, the Mistral-7B–based model achieves superior performance when approximately 7% of the training data (650 samples) is used. These findings indicate that our models require substantially less supervision to outperform existing fine-tuned strong baselines, trained using the full AEG training set. These experiments confirm the low-cost of our cross-tasking approach while maintaining strong performance gains over both zero-shot base LLMs, allowing us to mitigate current challenges in using LLMs for AG and AR.

6 Conclusion

We demonstrate that AG and AR can effectively leverage cross-tasking foundational skills

			Argument Generation					
			AEG (BLEU-4)			iDebate (BLEU)		
Llama-3-8B (ZS)			4.45			2.70		
			PEFT	FFT	Δ%	PEFT	FFT	Δ%
Cross-Task	ArgLLM_{MT}	1B	-5.4%	-3.4%	-2.0% ↓	-8.5%	+12.6%	-21.1% ↓
		3B	-4.9%	+3.1%	-8.0% ↓	+14.8%	+17.4%	-2.6% ↓
		8B	+16.4%	-8.5%	+24.9% ↑	+35.2%	+3.7%	+31.5% ↑
	ArgLLM_{MG}	1B	-7.9%	-3.6%	-4.3% ↓	+4.4%	+10.0%	-5.6% ↓
		3B	+8.5%	+3.6%	+4.9% ↑	+22.2%	+21.1%	+1.1% ↑
		8B	+21.3%	-0.4%	+21.7% ↑	+34.8%	+15.9%	+18.9% ↑
			Argument Retrieval					
			Arguana (NDCG@10)			**Touché (NDCG@10)**		
Llama-3-8B (ZS)			32.14			33.35		
			PEFT	FFT	Δ%	PEFT	FFT	Δ%
Cross-Task	ArgLLM_{MT}	1B	-1.8%	-43.06%	+41.3% ↑	-0.6%	-21.6%	+21.0% ↑
		3B	-80.3%	-82.8%	+2.5% ↑	-44.0%	-42.4%	-1.6% ↓
		8B	-13.1%	16.3%	+3.2% ↑	-8.1%	+11.2%	-19.3% ↓
	ArgLLM_{MG}	1B	-16.7%	-42.8%	+26.1% ↑	-6.2%	-26.5%	+20.3% ↑
		3B	-81.3%	-79.34%	-2.0% ↓	-44.8%	-43.7%	-1.1% ↓
		8B	+4.0%	-6.9%	+11.0% ↑	+11.8%	+17.0%	-5.2% ↓

Table 4: Percentage gain/loss of ArgLLM relative to the Llama-3-8B (ZS) baseline. PEFT and FFT report relative changes. Δ denotes the gap (PEFT% - FFT%).

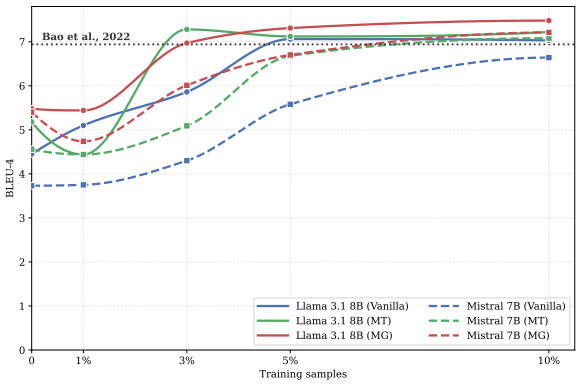


Figure 3: Performances on AEG dataset for Vanilla and ArgLLM models from Zero-Shot to Few-Shot Fine-Tuning. Dashed line indicates the performance for Bao et al. (2022).

that cover argument mining and argument quality assessment, either by multi-tasking or merging. Through *ArgLLM* framework, our empirical results show that cross-tasking significantly outperforms base LLMs in zero-shot, that it is robust to a range of cross-tasking strategies, and exhibits a good trade-off between computational cost and task performance. We believe that our findings can pave the way to impactful future research in NLA. Promising directions include: (1) designing NLA models that learn to compose atomic task-specific LLMs to achieve unseen complex tasks; (2) enhancing NLA-based applications (e.g., argument search engines) with inherent interpretability by allowing faithfulness of the outcomes based on the underlying atomic NLA tasks performed.

Limitations

Although we manually designed the templates used for dataset transformation, no manual nor full verification was performed on the transformed dataset,

which may propagate errors present in the original data or induce new ones during the transformation. Also, only one template was used per atomic-task dataset which maybe considered as limited when compared to the richness that human-generated MCQA may have. Although no verification was performed, some data contamination of the test set may exist as we used public available pre-trained models. However, note that we used the templates that changed the way that questions were presented to the LLMs which may reduce any data contamination issue. For example, in the case of ARL, we combined separated input text and added marking tokens to indicate specific content which may reduce the contamination effect, if any. Finally, due to a limited compute budget, we focus the evaluation on two small-sized LLMs. Although we used GPUs, our PEFT experiments can be run in single-GPU (H100) units while Full-Finetuning may need multiple-GPUs (up to 4 H100). Regarding the evaluation metrics on the complex task, we limited our experiments to standard metrics to the given datasets without exploring more recent setups such as LLM-as-a-judge. Last but not least, an in-depth analysis of the impact of levels of heterogeneity between task categories and across categories would have strengthened our conclusions, although this would have significantly increased the computational budget.

Acknowledgments

This work was supported by the AI Interdisciplinary Institute ANITI, funded by the French program AI Cluster ANR-23-IACL-0002 as well as the GUIDANCE project n°ANR-23-IAS1-0003 from the ANR - FRANCE (French National Research Agency).

References

Mohammad Yeghaneh Abkenar, Weixing Wang, Hendrik Graupner, and Manfred Stede. 2024. Assessing open-source large language models on argumentation mining subtasks. *arXiv preprint arXiv:2411.05639*.

AI@Meta. 2024. Llama 3 model card.

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. AEG: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project Debater APIs: Decomposing the AI grand challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. 2020. Overview of touché 2020: argument retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 384–395. Springer.

Victor Bursztyn, David Demeter, Doug Downey, and Larry Birnbaum. 2022. Learning to perform complex tasks through compositional fine-tuning of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1676–1686, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anwoy Chatterjee, Eshaan Tanwar, Subhabrata Dutta, and Tanmoy Chakraborty. 2024. Language models can exploit cross-task in-context learning for data-scarce novel tasks. In *Annual Meeting of the Association for Computational Linguistics*.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024a. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2024b. Skills-in-context: Unlocking compositionality in

large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13838–13890, Miami, Florida, USA. Association for Computational Linguistics.

Kaustubh Dhole, Kai Shu, and Eugene Agichtein. 2025. ConQRet: A new benchmark for fine-grained automatic evaluation of retrieval augmented computational argumentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5687–5713, Albuquerque, New Mexico. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Kaveh Eskandari Miandoab and Vasanth Sarathy. 2024. "let's argue both sides": Argument generation can force small models to utilize previously inaccessible reasoning capabilities. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 269–283, Miami, Florida, USA. Association for Computational Linguistics.

Amirhossein Farzam, Shashank Shekhar, Isaac Mehlhaff, and Marco Morucci. 2024. Multi-task learning improves performance in deep argument mining models. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 46–58, Bangkok, Thailand. Association for Computational Linguistics.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2024. Argumentative large language models for explainable and contestable decision-making.

Damián Furman, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore. Association for Computational Linguistics.

Debela Gemechu and Chris Reed. 2024. External knowledge-driven argument mining: Leveraging attention-enhanced multi-network models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3688–3709, Miami, Florida, USA. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can large language models perform relation-based argument mining? *arXiv preprint arXiv:2402.11243*.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276, Bangkok, Thailand. Association for Computational Linguistics.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5065–5069. International Joint Conferences on Artificial Intelligence Organization. Journal track.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XIX*, page 239–254, Berlin, Heidelberg. Springer-Verlag.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yohan Jo, Haneul Yoo, Jinyeong Bak, Alice H. Oh, Chris Reed, and Eduard H. Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. *ArXiv*, abs/2109.09057.

Wei-Yu Kao and An-Zi Yen. 2024. MAGIC: Multi-argument generation with self-refinement for domain

generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10891–10902, Torino, Italia. ELRA and ICCL.

Tushar Khot, Kyle Richardson, Daniel Khashabi, and Ashish Sabharwal. 2021. Hey ai, can you solve complex tasks by talking to agents? In *Association of Computational Linguistics (ACL) Findings*.

Johannes Kiesel, Çağrı Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaž Erjavec, Matthias Hagen, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Harrisen Scells, Moritz Wolter, Ines Zelch, Martin Potthast, and Benno Stein. 2025. Overview of touché 2025: Argumentation systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings*, page 486–508, Berlin, Heidelberg. Springer-Verlag.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. 43(3).

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore. Association for Computational Linguistics.

Zhiyue Liu, Jiahai Wang, and Zhenghong Li. 2021. Topic-to-essay generation with comprehensive knowledge enhancement. *ArXiv*, abs/2106.15142.

Xueguang Ma, Tommaso Teofili, and Jimmy Lin. 2023. Anserini gets dense retrieval: Integration of lucene's hnsw indexes. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5366–5370.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Luca Mouchel, Debjit Paul, Shaobo Cui, Robert West, Antoine Bosselut, and Boi Faltings. 2024. A logical fallacy-informed framework for argument generation. *arXiv preprint arXiv:2408.03618*.

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, et al. 2022. Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *International conference of the cross-language evaluation forum for european languages*, pages 495–520. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Joshua Robinson and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *International Conference on Learning Representations*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Maja Stahl, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth. 2025. ArgInstruct: Specialized instruction fine-tuning for computational argumentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11103–11127, Vienna, Austria. Association for Computational Linguistics.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. 2025. Model merging with SVD to tie the knots. In *The Thirteenth International Conference on Learning Representations*.

Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic evaluation of neural retrieval models on the touché 2020 argument retrieval subset of beir. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information*

Retrieval, SIGIR '24, page 1420–1430, New York, NY, USA. Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "pagerank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023a. Ties-merging: resolving interference when merging models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023b. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Qinyuan Ye. 2024. Cross-task generalization abilities of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 255–262, Mexico City, Mexico. Association for Computational Linguistics.

Kun Zhang, Jiali Zeng, Fandong Meng, Yuanzhuo Wang, Shiqi Sun, Long Bai, Huawei Shen, and Jie Zhou. 2024. Tree-of-reasoning question decomposition for complex question answering with large language models. AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yi Zhang, Sujay Kumar Jauhar, Julia Kiseleva, Ryen White, and Dan Roth. 2021. Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2726–2735, Online. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang 0002, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47.

A Argument MCQA instruction dataset

Here we present the text-based inputs to baselines and our models.

A.1 Full MCQA examples

For each atomic task, we designed a dataset specific transformation from an original text input to a MCQA-based input. Examples for each atomic task are presented in the Table 5.

A.2 Instructions used

Although MCQA allows an input normalization, each atomic task needs a task-specific instruction in order to inform the LLMs which option to pick. The details of these instructions are presented in Table 6.

B Implementation details

Models and training. All models are trained using the Hugging Face PEFT library (Mangrulkar et al., 2022), employing LoRA (Hu et al., 2022). We configure the LoRA adapters with a rank $r = 64$, a scaling factor $\alpha = 16$, and a dropout rate of 0.05. Following standard efficiency practices, no additional bias parameters are trained. LoRA adapters are applied to the q, k, v, and o projections, as well as the gate, up, and down projections, while all remaining model parameters are kept frozen. Training is performed for 5 epochs with automatic batch size selection, a learning rate of 2×10^{-4}, weight decay of 0.01, gradient accumulation of 4 steps, a warmup ratio of 0.03, and a cosine learning rate scheduler. Full fine-tuning experiments are conducted for 3 epochs using a batch size of 8, a learning rate of 1×10^{-5}, and weight decay of 0.01.

For model merging, we combine task-specific adapters using uniform interpolation weights ($\lambda = 0.33$), for each model across all merging methods and for TIES, we retain the top 80% of parameter deltas by magnitude, which we found empirically to work well and did not further tune (Wortsman et al., 2022; Yadav et al., 2023b). We implement merging using PEFT and MergeKit libraries.[5]

AG and AR considerations. Although the AG task does not require extra considerations, the AR task does. Indeed, we use the logits-based re-ranking technique proposed by Zhuang et al. (2024). It consists of giving the query and a list of retrieved documents as input to the first-stage model ranker and then re-ranking them based on the logit values of document positioning ids. This method requires a ranker that provides a small list of candidates and only requires one inference iteration by query. In our experiments, we use 20 documents for re-ranking as well as the BM25 and SPLADE (Formal et al., 2021) models as first-stage rankers. For comparative evaluation of AG and AR across datasets, we use BLEU-based metrics and NDCG-based metrics (in percentage), respectively.

[5] https://github.com/arcee-ai/mergekit

C Additional Experiments

C.1 Qualitative analysis

In order to access the individual contribution of each model trained on an atomic task, we analyzed the alignment between the output obtained by a cross-task model and atomic-task models for the following sample of the AEG dataset:

> "Topic: Mobile phones and the internet are very useful. However, it is rare for old people to use them. What ways could mobile phones and the internet be useful to old people? How does the old people to be encouraged using this new technology?"

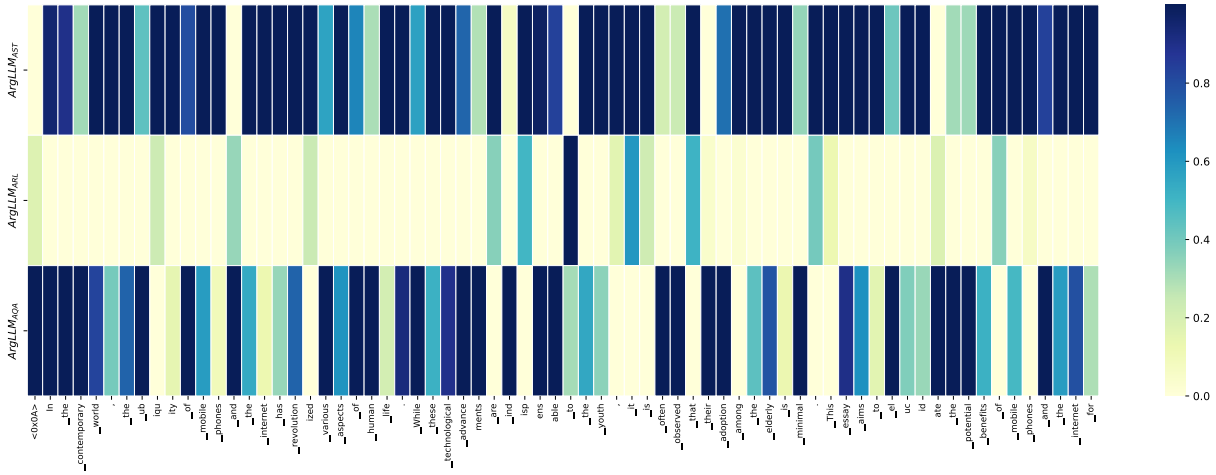
To do so, we used the outputs of cross-task models, ArgLLM$_{MT}$ and ArgLLM$_{MG}$, and assigned alignment value to each token by considering its ranking in a ordered list by logits extracted from each atomic model, namely ArgLLM$_{AST}$, ArgLLM$_{ARL}$, and ArgLLM$_{AQA}$. These values are used to build a heatmap at token level presented in Figure 4, where high values indicate a logit close to 1st position on the rank while low values indicate close to 1000th position, or even farther. From this figure, we observe that: (1) the ArgLLM$_{AST}$ is the model having the most high logits with both cross-task models outputs, which explains why 7 out of 8 ablation configurations (§ 5.3) performed worse without the structure component; (2) ArgLLM$_{MT}$ aligns with ArgLLM$_{ARL}$ more than ArgLLM$_{AQA}$, while ArgLLM$_{MG}$ follows a different pattern with a stronger alignment with ArgLLM$_{AQA}$. However, as mentioned in previous observation, both strongly align to ArgLLM$_{AST}$; (3) for each token there is a model with highly ranked logit, each token generated by ArgLLM$_{MT}$ or ArgLLM$_{MG}$ was triggered by either one or multiple atomic models (high logits at least in one model) which shows that the cross task models combine differently the knowledge of the atomic-models.

Task Category	Task	Dataset	text	question	options	answer
AST	US	Essay	topic: Computers have made life easier text: The advent of the computer is one of the results of the development of the advantaged technology. While some people advocate for the idea that computers have made life more complex and stressful, others support the idea that computers have made life easier and more convenient. In my point of view, computers brought convenience and easiness to our life since they enabled easier access to information and increased qualities of the communication. First, computers through internet made our access to information easier. Through computers we can collect appropriate data very quickly, store it in the hard disc as long as we want and use it when needed. Search engines such as google, yandex with many features relevant to the characteristics of the information needed enable search of information in few minutes. Social networks such as facebook, twitter have changed traditional sources of information and decreasing the monopolization of the information by the governments. Hence, with the use of computers we spend less time in search for information, and can get it from different sources. Second, in a globalized world that we live computers made communications faster, visual and cheaper. With the use of email our messages can be delivered to any person in any part of the world in few minutes. In addition, it is convenient because you do not need to mile to the post office, but just click the mouse of your computer from the convenience of your arm chair in order to reach your friends, family members in any part of the world. Internet facilities such as msn, skype made visual communication possible. For example, via skype people not just communicate with their family members and friends, but also make interviews for employment, enrolment into universities. In result, communication has become cheaper, faster and visual. Finally, I believe that computers have made life easier and more convenient in the ways that it broadened the sources of information and made access more quick, a long with the improving quality of the communication in many ways.	<topic>Computers have made life easier</topic> <essay>The advent of the computer is one of the results of the development of the advantaged technology. While some people advocate for the idea that computers have made life more complex and stressful, others support the idea that computers have made life easier and more convenient. In my point of view, computers brought convenience and easiness to our life since <s>they enabled easier access to information and increased qualities of the communication</s>. First, computers through internet made our access to information easier. Through computers we can collect appropriate data very quickly, store it in the hard disc as long as we want and use it when needed. Search engines such as google, yandex with many features relevant to the characteristics of the information needed enable search of information in few minutes. Social networks such as facebook, twitter have changed traditional sources of information and decreasing the monopolization of the information by the governments. Hence, with the use of computers we spend less time in search for information, and can get it from different sources. Second, in a globalized world that we live computers made communications faster, visual and cheaper. With the use of email our messages can be delivered to any person in any part of the world in few minutes. In addition, it is convenient because you do not need to mile to the post office, but just click the mouse of your computer from the convenience of your arm chair in order to reach your friends, family members in any part of the world. Internet facilities such as msn, skype made visual communication possible. For example, via skype people not just communicate with their family members and friends, but also make interviews for employment, enrolment into universities. In result, communication has become cheaper, faster and visual. Finally, I believe that computers have made life easier and more convenient in the ways that it broadened the sources of information and made access more quick, a long with the improving quality of the communication in many ways.</essay> What type of argument is the segment marked with <s>in this essay on <topic>?	A. MajorClaim B. Premise C. Claim D. None of the other options	C. Claim
	CD	CheckThat2022	Native American communities have been hit hard by COVID-19. But thanks to the Indian Health Service and strong partnerships with Tribal governments, organizations, and urban Indian groups, more than 500,000 vaccines have already been administered with more on the way.	<tweet>Native American communities have been hit hard by COVID-19. But thanks to the Indian Health Service and strong partnerships with Tribal governments, organizations, and urban Indian groups, more than 500,000 vaccines have already been administered with more on the way.</tweet> Does the <tweet>contain a factual and verifiable claim related to COVID-19?	A. Yes B. No	A. Yes
ARL	SC	ibm	Topic: This house would build hydroelectric dams</topic> Claim: As time progresses, renewable energy generally gets cheaper,[ref][ref] while fossil fuels generally get more expensive</claim>	<topic>This house would build hydroelectric dams</topic><claim>As time progresses, renewable energy generally gets cheaper,[ref][ref] while fossil fuels generally get more expensive</claim> What is the stance of the claim on the topic?	A. Favor B. Against	A. Favor
	SA	essay2	Topic:Traditional games or modern games in developing children's skills Essay:Games have played a key role in children's growth, especially in terms of their abilities. With technological advances, children have more access to modern games currently. In such case, the relative importance of traditional games and modern games in children's developments of skills has become a frequent topic of discussion. For me, I believe parents and educators should attach more importance to traditional games. It is true that modern games may be, to some extent, beneficial for children to foster some skills, such as computer skills or the capacity to keep up with the latest trend. This is because children have to be proficient at computers and the Internet if they want to join online games, which, in fact, helps children acquire a particularly powerful skill at work in future. At the same time, children have the chance to experience the state-of-the-art technology, raising their awareness of innovation rather than stay conservative. However, I think traditional games are still indispensible in children's learning process, even much more essential than modern games, especially in modern society. One primary merit of traditional games is that they foster children's communication skills. <p1>Unlike most modern games which focus on the interactions between children and machines, traditional games provide a relaxing and enjoyable atmosphere where children can chat, laugh and cooperate face to face</p1>. As a result, <p2>communicating with a variety of people will not be an issue for these children any more</p2>. Furthermore, it is the educational functions traditional games hold that keep them alive today. In fact, these traditional games were elaborately devised by educators and have been proven effective in improving children's skills on different aspects in previous teaching practices. By contrast, modern games are developed by game companies for the purpose of profits. Therefore, there is a risk that children may be exposed to unhealthy contents, such as violence or pornography, arranged in the games by such companies to secure financial survival. Thus, I would conclude that traditional games should be, by no means, ignored by parents and teachers with the advent of modern games. Only through traditional games can children be ensured a positive and healthy skills learning process.	<topic>Traditional games or modern games in developing children's skills</topic> <essay>Games have played a key role in children's growth, especially in terms of their abilities. With technological advances, children have more access to modern games currently. In such case, the relative importance of traditional games and modern games in children's developments of skills has become a frequent topic of discussion. For me, I believe parents and educators should attach more importance to traditional games. It is true that modern games may be, to some extent, beneficial for children to foster some skills, such as computer skills or the capacity to keep up with the latest trend. This is because children have to be proficient at computers and the Internet if they want to join online games, which, in fact, helps children acquire a particularly powerful skill at work in future. At the same time, children have the chance to experience the state-of-the-art technology, raising their awareness of innovation rather than stay conservative. However, I think traditional games are still indispensible in children's learning process, even much more essential than modern games, especially in modern society. One primary merit of traditional games is that they foster children's communication skills. <p1>Unlike most modern games which focus on the interactions between children and machines, traditional games provide a relaxing and enjoyable atmosphere where children can chat, laugh and cooperate face to face</p1>. As a result, <p2>communicating with a variety of people will not be an issue for these children any more</p2>. Furthermore, it is the educational functions traditional games hold that keep them alive today. In fact, these traditional games were elaborately devised by educators and have been proven effective in improving children's skills on different aspects in previous teaching practices. By contrast, modern games are developed by game companies for the purpose of profits. Therefore, there is a risk that children may be exposed to unhealthy contents, such as violence or pornography, arranged in the games by such companies to secure financial survival. Thus, I would conclude that traditional games should be, by no means, ignored by parents and teachers with the advent of modern games. Only through traditional games can children be ensured a positive and healthy skills learning process.</essay> What is the relation between the two premises enclosed in <p1>and <p2>tags in the essay?	A. Support B. Attack	A. Support
AQA	OA	dagstuhl	Topic: Gay marriage right or wrong Stance: Allowing gay marriage is right Argument: How can anyone say that gay marriage is wrong. it is a personal choice that is made from personal beliefs so who are we to say that gay couples do not have the right to enjoy all of the benefits that straight couples do?	<topic>Gay marriage right or wrong</topic> <argument>How can anyone say that gay marriage is wrong. it is a personal choice that is made from personal beliefs so who are we to say that gay couples do not have the right to enjoy all of the benefits that straight couples do</argument> How would you rate the acceptability of the author's argument toward the topic on the stance 'Allowing gay marriage is right'?	A. Low B. High C. Medium	C. Medium
	AU	WebisArgRank17	Conclusion: We lack any sense of the common good.	<conclusion>We lack any sense of the common good.</conclusion> Which argument best supports the conclusion?	A. If an attack causes more death injury and phsycholoigical and economic damage than others in a series of attacks it is by definition the deadliest attack in the series B. The only thing we hold in common is something abstract, which is money.	B. The only thing we hold in common is something abstract, which is money.

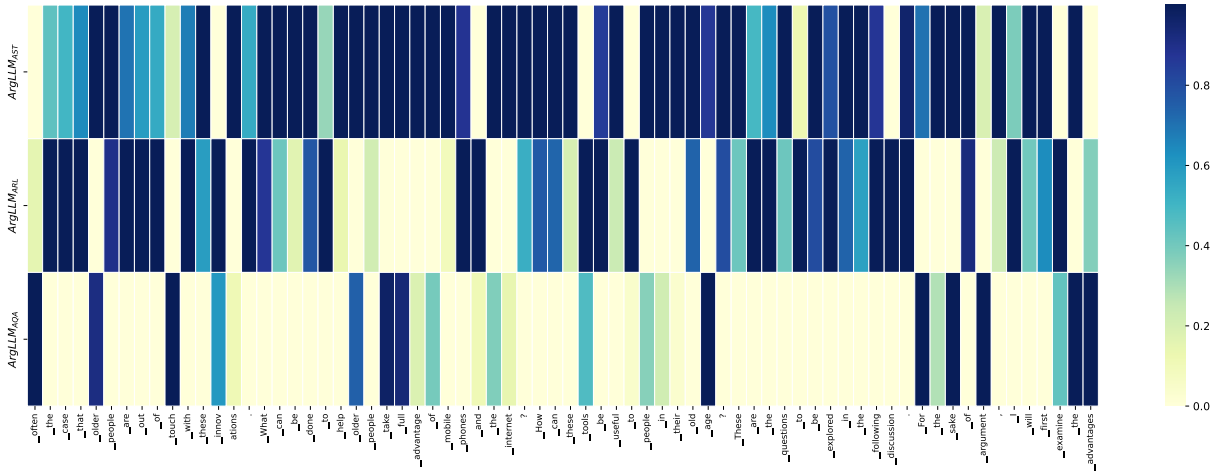
Table 5: Example of transformed MCQA for each dataset in *ArgLLM* framework.. "text" correspond to orignal text found in the dataset while "question" is our transformed question.

Category	Task/Dataset	System	User
AST	Unit Segmentation (US)	You are given a sentence from a student's argumentative essay. Your goal is to determine the argumentative role of the sentence. Possible roles are: Major Claim, Claim, Premise, or None of the other options.	Segment:\n{question} Options:\n{options} Answer: {answer}
	Claim Detection (CD)	You are given a sentence from an argumentative text. Determine whether this sentence contains a factual claim. Answer with the correct option.	Sentence:\n{question} Options:\n{options} Answer: {answer}
ARL	Stance Classification (SC)	You are a helpful assistant specialized in stance classification. For each question, you are given a claim and a topic, followed by two possible answers labeled A and B. Your task is to determine whether the claim supports or opposes the belief expressed in the topic, based on the provided choices. Respond with the letter of the correct option followed by its full text.	Question:\n{question} Options:\n{options} Stance: {answer}
	Support/ Attack Classification (SA)	Your task is to determine the relationship between two argumentative units, which are marked with tags in a given text. These units can either support or attack each other. You will be given a multiple-choice question, and the possible answers are labeled with the letters A and B. Provide your answer by selecting only the correct options.	Question:\n{question} Options:\n{options} Relation: {answer}
AQA	Argument Reuse (AU)	You are a precise assistant for evaluating arguments. You will be given a conclusion and two arguments labeled A and B. Your task is to choose which argument better supports the conclusion. Respond with ONLY one uppercase letter: 'A' or 'B'. Do not explain your choice.	Conclusion: {question} Options:\n{options} Label:{answer}
	Objective Acceptability (OA)	You are an expert assistant tasked with evaluating the acceptability of a comment from an online debate forum on a specific topic. Acceptability means that at least one member of the target audience would find the comment relevant and well-expressed. You will be given a multiple-choice question with labeled options, and only one option is correct. Respond with only the correct option.	Question:\n{question} Options:\n{options} Acceptability: {answer}

Table 6: Instructions used during training in each dataset of the atomic tasks. Strings into "{}" represent variables that will be filled with values, such as in examples present in Table 5. System and User are inputs for the LLM.



(a) ArgLLM$_{MG}$'s output.



(b) ArgLLM$_{MT}$'s output.

Figure 4: Comparative analysis of atomic models (ArgLLM$_{AST}$, ArgLLM$_{ARL}$ and ArgLLM$_{AQA}$) influence on next-token prediction, for generated output from ArgLLM$_{MT}$ and ArgLLM$_{MT}$

C.2 Atomic-task results

We evaluated the performance of our models versus a zero-shot baseline. Results show that our models are able to capture correctly the tasks of AST and ARL, as well as the Argument Reuse of the AQA task. However, both models were unable to improve a zero-shot configuration in the Objective Acceptability task. Detailed results are presented in Table 7.

Task	Subtask	Model	Zero-Shot Prompting	Fine Tuning
AST	Unit Segmentation	Mistral-7B	29.2	**84.4**
		Llama-3-8B	29.4	**85.0**
ARL	Claim Detection	Mistral-7B	65.6	**79.9**
		Llama-3-8B	67.8	**77.9**
	Stance Classification	Mistral-7B	60.97	**88.63**
		Llama-3-8B	63.47	**85.31**
AQA	Support/Attack Relation	Mistral-7B	90.11	**94.31**
		Llama-3-8B	81.33	**94.81**
	Argument Reuse	Mistral-7B	66.67	**80.56**
		Llama-3-8B	72.22	**83.33**
	Objective Acceptability	Mistral-7B	**32.81**	29.69
		Llama-3-8B	**45.31**	43.75

Table 7: Comparison between zero-shot and fine-tuned models on atomic tasks.

C.3 Argument Generation

Detailed performances on the argument generation are presented in Table 8. For each dataset (AEG and iDebate), we present officials metrics of the models trained in each task category ($ArgLLM_{AST}$, $ArgLLM_{ARL}$, and $ArgLLM_{AQA}$) and our cross-tasking strategies $ArgLLM_{MT}$ and $ArgLLM_{MG}$. Note that this are the same experiments than in Table 2 but extra

metrics were added for the sake of transparency in our results.

C.4 Argument Retrieval

Detailed performances on the argument retrieval are presented in Table 9. For each dataset (Arguana and Touché2020), we present standard retrieval metrics of the models trained in each task category ($ArgLLM_{AST}$, $ArgLLM_{ARL}$, and $ArgLLM_{AQA}$) and and our cross-tasking strategies $ArgLLM_{MT}$ and $ArgLLM_{MG}$. Note that this are the same experiments than in Table 2 but extra metrics were added for the sake of transparency in our results.

C.5 Comparative analysis

Comparative analysis of the *ArgLLM* models vs task-specific state of the art. Note that more recent works on AR datasets, such as (Ma et al., 2023), do not outperform (Thakur et al., 2021) results.

C.6 Impact of instruction format

Previous work showed that prompt templates are key in performing complex tasks, with prompt length being one of the most impactful features (He et al., 2024). Specifically, since we aim to perform AG and AR in zero-shot settings, we evaluate the effect of prompt length by truncation. As an example of the truncation, our original prompts for iDebate and Touché are:

> *"Given a query {query}, which of the following passages is the most relevant to the query? Favor passages with high-quality, well-structured argumentative content"*
> *"Here are reasons supporting a claim. Rephrase them into one short sentence, using the same vocabulary where possible. Avoid generalizations."*

The respective truncated versions are:

> *"Given a query {query}, which of the following passages is the most relevant to the query?"*
> *"Rephrase following text into one short sentence"*

Results of the performances of Arg-LLM$_{MT}$ and the zero-shot models for Mistral-7B are presented in Figure 5. We can see that both prompt template versions have a positive impact on the performance of both AG and AR compared to their counterpart vanilla LLMs. Besides, we can observe that for the AR task as well as the AEG dataset of the AG task, the results show that our prompt choice is adequate for both the trained atomic task models and the vanilla ones. However, the iDebate dataset from the complex AG task behaves differently. In this case, the truncated version of the prompt template outperforms our original prompt, suggesting that a prompt tuning strategy may further improve upon our results after training on this dataset.

C.7 Comparative analysis: PEFT vs Full FineTuning

Raw values of the comparative analysis between different model sizes using PEFT and Full FineTuning for *ArgLLM* are presented in Table 11. Note that this values where used in the percentage analysis included in Table 4.

		Argument Generation									
		AEG							iDebate		
		Dist-3	**Dist-4**	**Nov-1**	**Nov-2**	**Rep-3**	**Rep-4**	**BLEU-4**	**BLEU**	**METEOR**	**ROUGE SU-4**
ZeroShot	Mistral-7B-Instruct-v0.3	62.75	79.78	**78.80**	**95.63**	8.23	4.30	3.62	2.68	21.58	5.86
	Llama-3-8B-Instruct	51.15	68.71	75.32	94.49	21.30	14.98	4.46	2.70	**23.06**	6.14
$ArgLLM_{AST}$	Mistral-7B-Instruct-v0.3	64.73	81.85	74.21	94.26	8.40	4.34	4.60	4.10	18.41	7.13
	Llama-3-8B-Instruct	59.25	78.30	72.23	93.78	16.13	10.08	4.68	3.50	20.01	6.30
$ArgLLM_{ARL}$	Mistral-7B-Instruct-v0.3	**67.18**	**83.75**	74.43	94.34	7.17	3.59	4.50	**4.30**	18.83	**7.51**
	Llama-3-8B-Instruct	59.65	77.82	71.24	93.54	21.32	**15.13**	4.79	3.40	15.95	6.04
$ArgLLM_{AQA}$	Mistral-7B-Instruct-v0.3	62.65	79.79	78.73	95.59	8.63	4.51	3.64	3.64	20.70	6.61
	Llama-3-8B-Instruct	50.82	68.41	75.45	94.54	**21.39**	14.96	4.38	3.83	19.39	6.82
$ArgLLM_{MT}$	Mistral-7B-Instruct-v0.3	66.16	83.18	73.68	94.29	8.67	4.65	4.56	2.77	11.96	3.24
	Llama-3-8B-Instruct	57.74	76.31	69.19	93.05	20.87	14.14	5.18	3.65	16.16	6.48
$ArgLLM_{MG}$	Mistral-7B-Instruct-v0.3	61.95	79.32	69.88	93.04	16.40	10.74	5.40	3.80	21.57	7.00
	Llama-3-8B-Instruct	59.94	77.94	72.57	93.54	11.64	6.81	**5.48**	3.64	20.11	6.68

Table 8: Argument generation results on AEG and iDebate.

		Argument Retrieval			
		Arguana		Touché2020	
		NDCG@10	**Recall@10**	**NDCG@10**	**Recall@10**
Zero-shot	Mistral-7B-Instruct-v0.3	33.08	70.55	36.69	21.62
	Llama-3-8B-Instruct	32.14	71.62	33.35	18.44
$ArgLLM_{AST}$	Mistral-7B-Instruct-v0.3	35.01	72.04	35.37	21.07
	Llama-3-8B-Instruct	34.80	**74.32**	29.67	19.42
$ArgLLM_{ARL}$	Mistral-7B-Instruct-v0.3	36.71	72.47	37.60	22.87
	Llama-3-8B-Instruct	22.78	53.05	36.34	22.23
$ArgLLM_{AQA}$	Mistral-7B-Instruct-v0.3	33.09	70.62	39.30	23.29
	Llama-3-8B-Instruct	32.13	71.55	32.42	18.73
$ArgLLM_{MT}$	Mistral-7B-Instruct-v0.3	**39.26**	71.69	**43.89**	**25.58**
	Llama-3-8B-Instruct	27.94	61.09	29.51	17.86
$ArgLLM_{MG}$	Mistral-7B-Instruct-v0.3	37.41	71.55	38.20	22.51
	Llama-3-8B-Instruct	33.43	68.71	37.27	22.98

Table 9: Argumentation retrieval results (NDCG@10 and Recall@10) for Arguana and Touché2020 datasets.

Complex Task	**Dataset (source)**	**Metric**	$ArgLLM_{MT}$	$ArgLLM_{MG}$	**Task specific SOTA**
Argument Generation	AEG	BLUE-4	5.18	5.48	6.94 (Bao et al., 2022)
	iDebate	BLUE	3.65	3.80	25.84 (Wang and Ling, 2016)
Argument Retrieval	Arguana	NDCG@10	39.26	37.41	49.3 (Thakur et al., 2021)
	Touché		43.89	38.20	36.7 (Thakur et al., 2021)

Table 10: Comparative analysis of the *ArgLLM* models vs task-specific state of the art. Note that more recent works on AR datasets, such as (Ma et al., 2023), do not outperform (Thakur et al., 2021) results.

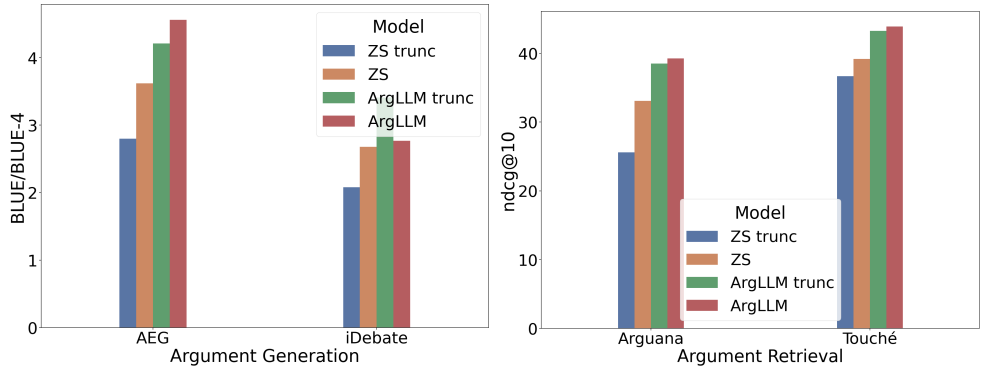


Figure 5: Impact of prompt templates on AG and AR performance using zero-shot (ZS) and ArgLLM$_{MT}$, with and without template truncation.

			Argument Generation				Argument Retrieval			
			AEG		**iDebate**		**Arguana**		**Touché**	
			BLEU-4		BLEU		NDCG@10			
			PEFT	FFT	PEFT	FFT	PEFT	FFT	PEFT	FFT
ZS		Llama-3-1B	3.68		1.65		24.26		26.12	
		Llama-3-3B	4.04		1.91		6.74		18.71	
		Llama-3-8B	4.45		2.70		32.14		33.35	
Cross-Task	ArgLLM_{MT}	Llama-3-1B	4.21	4.30	2.47	3.04	31.56	18.30	33.15	26.16
		Llama-3-3B	4.23	4.59	3.10	3.17	6.33	5.52	18.69	19.21
		Llama-3-8B	5.18	4.07	3.65	2.80	27.94	26.91	30.66	37.09
	ArgLLM_{MG}	Llama-3-1B	4.10	4.29	2.82	2.97	26.78	18.38	31.29	24.52
		Llama-3-3B	4.83	4.61	3.30	3.27	6.00	6.64	18.40	18.78
		Llama-3-8B	**5.40**	4.43	**3.64**	3.13	**33.43**	29.91	37.27	**39.01**

Table 11: Comparison between different model sizes with both PEFT and Full FineTuning for ArgLLM.