Argument Mining in BioMedicine: Zero-Shot, In-Context Learning and Fine-tuning with LLMs

Jérémie Cabessa[1,*,†], Hugo Hernault[2,†] and Umer Mushtaq[3,†]

[1]David Lab, University of Versailles Saint-Quentin (UVSQ) – University of Paris-Saclay, 78000 Versailles, France
Institute of Computer Science of the Czech Academy of Sciences, 18207 Prague 8, Czech Republic
[2]Playtika Ltd., CH-1003 Lausanne, Switzerland
[3]Laboratoire Informatique, Image, Interaction (L3i), University of La Rochelle, 17042 La Rochelle, France

Abstract

Argument Mining (AM) aims to extract the complex argumentative structure of a text and Argument Type Classification (ATC) is an essential sub-task of AM. Large Language Models (LLMs) have shown impressive capabilities in most NLP tasks and beyond. However, fine-tuning LLMs can be challenging. In-Context Learning (ICL) has been suggested as a bridging paradigm between training-free and fine-tuning settings for LLMs. In ICL, an LLM is conditioned to solve tasks using a few solved demonstration examples included in its prompt. We focuse on AM in the biomedical AbstRCT dataset. We address ATC using quantized and unquantized LLaMA-3 models through zero-shot learning, in-context learning, and fine-tuning approaches. We introduce a novel ICL strategy that combines kNN-based example selection with majority vote ensembling, along with a well-designed fine-tuning strategy for ATC. In zero-shot setting, we show that LLaMA-3 fails to achieve acceptable classification results, suggesting the need for additional training modalities. However, in our ICL training-free setting, LLaMA-3 can leverage relevant information from only a few demonstration examples to achieve very competitive results. Finally, in our fine-tuning setting, LLaMA-3 achieves state-of-the-art performance on ATC task in AbstRCT dataset.

Keywords

Argument Mining, NLP, LLMs, LLaMA-3, Zero-Shot Learning, In-Context Learning, Fine-tuning, Ensembling

1. Introduction

Argument Mining (AM) focuses on extracting the underlying argumentative and discursive structure from raw text [1]. Argument Type Classification (ATC), which involves classifying argumentative units in text according to their argumentative roles, is the crucial sub-task in AM. Research has shown that the argumentative role of a unit cannot be inferred solely for its text: additional structural and contextual information is needed [2]. This additional information can be incorporated via feature engineering [2], memory-enabled neural architectures [3, 4] or LLM-based hybrid methods [5, 6].

Large Language Models (LLMs) have become ubiquitous in deep learning and have shown impressive capabilities in most NLP tasks [7]. In the main, LLMs are used in two distinct settings: (i) training-free, where the pre-trained LLM is used for inference without any parameter adjustment, and (ii) fine-tuning, where the parameters of the LLM are updated through supervised training to enable transfer learning on a downstream task. Zero-shot learning refers to the training-free approach where a pre-trained LLM is prompted to solve tasks on completely unseen data samples.

Recently, In-Context Learning (ICL) has been proposed as a bridging paradigm between the training-free and fine-tuning settings. ICL is a prompt engineering technique whereby an LLM is conditioned to solve tasks by means of a few solved demonstration examples included as part of its input prompt [8]. Generally, the input prompt includes task instructions, the current input sample to be solved as well as several solved input-output pair examples. In this way, ICL maintains the training-free posture (parameters frozen) of the LLM while at the same time providing it with some supervision through demonstration examples. It also enables direct incorporation of selected features inside the prompt template, thereby obviating the need for architecture customization. Creative ICL strategies combining kNN-based examples selection, generated chain-of-thought (CoT) prompting, and majority vote ensembling have been proposed and shown to outperform fine-tuning approaches [9, 10, 11, 12]. In the main, kNN-based examples selection optimizes the process of learning from few examples and ensembling increases the robustness of the predictions [13, 9, 11].

This work focuses on AM in the biomedical AbstRCT dataset [14]. More specifically, we address the ATC task using quantized and unquantized LLaMA-3 models, among the most capable openly available LLMs (cf. leaderboard), through zero-shot learning, in-context learning,

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†The authors contributed equally.
jeremie.cabessa@uvsq.fr (J. Cabessa); hugoh@playtika.com (H. Hernault); umer.mushtaq@univ-lr.fr (U. Mushtaq)
0000-0002-5394-5249 (J. Cabessa); 0009-0003-4403-6048 (U. Mushtaq)

and fine-tuning approaches. Our contributions are as follows:

- In zero-shot learning setting, we show that LLaMA-3 fails to achieve acceptable classification results, suggesting the need for implementing additional training modalities.
- We introduce a novel ICL strategy that combines kNN-based example selection with majority vote ensembling. In this training-free setting, LLaMA-3 can leverage relevant information from only a few demonstration examples to achieve very competitive results.
- We further experiment with fine-tuning strategy for LLaMA-3. In this setting, we achieve state-of-the-art performance on the ATC task for AbstRCT dataset.

Our code is freely available on GitHub.

2. Related Works

In early works, Argument Mining has been approached using both classical algorithms such as SVM [15, 2, 16, 17] as well as recurrent neural network models such as BiLSTMs [18, 19, 4]. Transformer-based models, such as BERT [20], have also been utilized for AM, including multi-scale argument modelling and customized feature-injected BERT-based models [21, 22, 23, 5, 6, 24, 25]. AM in the biomedical AbstRCT dataset has been approached using LSTMs [26, 27], sequential transfer learning [28] as well as transformer-based models [29, 30, 31].

More recently, AM sub-tasks have been modeled as text generation tasks using LLMs. For the Argument Type Classification (ATC) sub-task, this approach involves using a prompt template to generate the corresponding class of an argument component. This method has been applied to various AM use-cases, such as podcast transcripts and legal documents [32, 33, 34]. The latest approach in this 'AM using LLM text generation' direction involves a prompt that includes the argument component as the query and the complete text as the context, to output the class of the argument component using a generative model [35]. In this study, the three AM sub-tasks are modeled using the Persuasive Essays (PE) and AbstRCT datasets.

In contrast to the fine-tuning approach, a relevant training-free ICL prompting strategy for LLMs has been proposed [9, 11]. This strategy combines kNN-based example selection, generated chain-of-thought prompting, and majority vote ensembling for few-shot classification. Interestingly, the ICL strategy outperforms the fine-tuning approach on the datasets used in the study.

Our work sits at the intersection of zero-shot learning, in-context learning and fine-tuning. We implement and compare the performance of the latest openly available LLMs using these three approaches for AM on the AbstRCT dataset.

3. Methodology

3.1. Datasets

We consider the AbstRCT dataset which consists of abstracts of 650 Randomized Controlled Trials selected from the biomedical database PUBMed [14]. For AbstRCt dataset, the Neoplasm train set (Neo-train) consists of 350 abstracts whereas the three Neoplasm, Glaucoma and Mixed tests sets (Neo-test, Gla-test and Mix-test, respectively) consist of 100 abstracts each. The statistics of AbstRCT dataset are given in Table 1. The argument type classification (ATC) task consists of predicting the type of each argument component (AC) as 'Major Claim', 'Claim' or 'Premise'. Following previous approaches, we combine the 'Major Claim' and 'Claim' classes into a single class 'Claim'.

Dataset Split	Abstracts	ACs
Neo-train	350	2,291
Neo-test	100	691
Gla-test	100	615
Mix-test	100	609

Table 1
AbstRCT dataset statistics.

An sample of the AbstRCT dataset is provided below. The argument components (ACs) and their corresponding classes are indicated by bold tags.

<**AC1: Major Claim**>A combination of mitoxantrone plus prednisone is preferable to prednisone alone for reduction of pain in men with metastatic, hormone-resistant, prostate cancer.</**AC1**> The purpose of this study was to assess the effects of these treatments on health-related quality of life (HQL). Men with metastatic prostate cancer (n = 161) were randomized to receive either daily prednisone alone or mitoxantrone (every 3 weeks) plus prednisone. Those who received prednisone alone could have mitoxantrone added after 6 weeks if there was no improvement in pain. HQL was assessed before treatment initiation and then every 3 weeks using the European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire C30 (EORTC QLQ-C30) and the Quality of Life Module-Prostate 14 (QOLM-P14), a trial-specific module developed for this study. An intent-to-treat analysis was used to determine the mean duration of HQL improvement and differences in improvement duration between groups of patients. <**AC2: Premise**> At 6 weeks, both groups showed improvement in several HQL domains</**AC2**>, and <**AC3: Premise**>only physical functioning and pain were better in the mitoxantrone-plus-prednisone group than in the prednisone-alone group</**AC3**>. <**AC4: Premise**>After 6 weeks, patients taking prednisone showed no improvement in HQL scores, whereas those taking mitoxantrone plus prednisone showed significant improvements in global quality of life (P =.009), four functioning domains, and nine symptoms (.001 < P <. 01)</**AC4**>, and <**AC5: Premise**>the improvement (> 10 units on a scale of 0

to100) lasted longer than in the prednisone-alone group (.004 < P <.05)</**AC5**>. <**AC6: Premise**>The addition of mitoxantrone to prednisone after failure of prednisone alone was associated with improvements in pain, pain impact, pain relief, insomnia, and global quality of life (.001 < P <.003).</AC6> <**AC7: Claim**>Treatment with mitoxantrone plus prednisone was associated with greater and longer-lasting improvement in several HQL domains and symptoms than treatment with prednisone alone.</**AC7**>

3.2. Zero-Shot Learning (ZSL) and In-Context Learning (ICL)

Zero-shot learning (ZSL) is the paradigm where the LLM is asked to solve a downstream task without receiving any specific solved examples in the prompt. By contrast, *in-context learning (ICL)* refers to the emergent ability of LLMs to solve a downstream task based on a few demonstration examples given in the prompt as contextual information [8]. As the major advantage, ZSL and ICL paradigms do not require any fine-tuning of the model's parameters (i.e. training-free framework).

Formally, let x be a query input text and $C = [I; t(x_{i_1}, y_{i_1}); \ldots; t(x_{i_k}, y_{i_k})]$ be a context composed of instructions I concatenated with input-output pairs (x_j, y_{i_j}) in text format, where $X = \{x_1, x_2, \ldots\}$ and $Y = \{y_1, \ldots, y_k\}$ are the sets of possible input and outputs, respectively. The ZSL and ICL paradigms correspond to the cases where $k = 0$ and $k > 0$, respectively. For input x, the LLM $\mathcal{M}$ predicts the output $\hat{y}$ such that

$$\hat{y} = \arg\max_{y_i \in Y} P_{\mathcal{M}}(y_i \mid C; x)\,,$$

where $P_{\mathcal{M}}(y_i \mid C; x)$ is the probability that $\mathcal{M}$ generates y_i when C and x are given as prompt. The main rationale behind ZSL and ICL is that the consideration of a well-chosen context C increases the probability of $\mathcal{M}$ predicting the correct answer y for input x, i.e., that $P_{\mathcal{M}}(y \mid C; x) > P_{\mathcal{M}}(y \mid x)$.

We consider a 2-step ICL strategy for argument type classification (ATC) inspired by a recent study [9] (see Figure 1). More precisely, let A be an abstract containing argument components (ACs) $c_1, \ldots, c_m$ with corresponding true classes $y_1, \ldots, y_m$, where each $y_i \in \{\text{Claim}, \text{Premise}\}$. Given the ACs $c_1, \ldots, c_m$ in the prompt, the LLM generates the corresponding class predictions $\hat{y}_1, \ldots, \hat{y}_m$ as follows:

(1) **kNN-based examples selection ($k = 3, 5$):** First, $2k$ neighboring abstracts $A_1, \ldots, A_{2k}$ of A are selected according to the following similarity measure. For any abstract A_i, let the signature of A_i be the embedding of the first sentence of A_i using the BioBERT model. The abstracts $A_1, \ldots, A_{2k}$ are the ones whose signatures are the closest, with respect to cosine similarity, to the signature of A. Then, k abstracts, $A_{i_1}, \ldots, A_{i_k}$, are randomly chosen from $A_1, \ldots, A_{2k}$. Afterwards, a prompt containing all the ACs and their corresponding classes in these k abstracts is constructed (**kNN**). Finally, the LLM predicts the classes $\hat{y}_1, \ldots, \hat{y}_m$ of $c_1, \ldots, c_m$ on the basis of on this prompt.

(2) **n-Ensembling ($n = 3, 5$):** The kNN-based examples selection step, which involves randomness, is repeated n times (**nEns**), leading to a set of n sequences of class predictions $\{(\hat{y}_{i,1}, \ldots, \hat{y}_{i,m}) : i = 1, \ldots n\}$. The final class predictions $\hat{y}_1, \ldots, \hat{y}_m$ of $c_1, \ldots, c_m$ are obtained by applying a component wise majority vote to the n predictions sequences.

The kNN-based example selection optimizes learning from few examples by selecting samples most similar to the current instance, rather than choosing them randomly. The ensembling step increases prediction robustness by selecting the most frequent predictions. Note that the relevance of the ensembling step relies on the random selection in the kNN step. This randomness ensures that same predictions are not always produced, allowing for majority voting and thereby increasing robustness.

To aid the LLM in generating predictions, additional task-specific information is typically included in the prompt. For example, definitions of the 'Claim' and 'Premise' classes, along with their statistics in the Neotrain set, can be incorporated in the prompt (**info**). Moreover, in addition to the ACs $c_1, \ldots, c_m$ whose class are to be predicted, the abstract text from which these ACs originate can be included in the prompt (**abstract**). According to this ICL strategy, the classes $\hat{y}_1, \ldots, \hat{y}_m$ of $c_1, \ldots, c_m$ are predicted all-at-once (see Figure 1). Therefore, a prompt of the form 'info + abstract + 3NN + 3Ens' (see Table 3) indicates that the argument components (ACs) of the abstract are predicted all-at-once, by incorporating additional information and the entire abstract text as contextual cues in the prompt, and employing the ICL strategy with 3NN-based example selection and 3-ensembling. A similar ICL strategy, where the classes $\hat{y}_1, \ldots, \hat{y}_m$ are inferred one-by-one (i.e., each model inference leads to a single prediction $\hat{y}_j$), has been considered but shown to be significantly less efficient. Due to space constraints, the latter results are omitted in this work.

3.3. Fine-tuning

Fine-tuning (FT) refers to the process of further training a pre-trained LLM on a downstream task. Previous studies indicate that relying solely on the text of an argument component is insufficient for predicting its argumentative class; additional contextual information is essential for achieving competitive classification accuracy [2, 5, 6]. Therefore, we propose a fine-tuning strategy that models the ATC task at the document level. Specifically, we incorporate task-specific information into each training

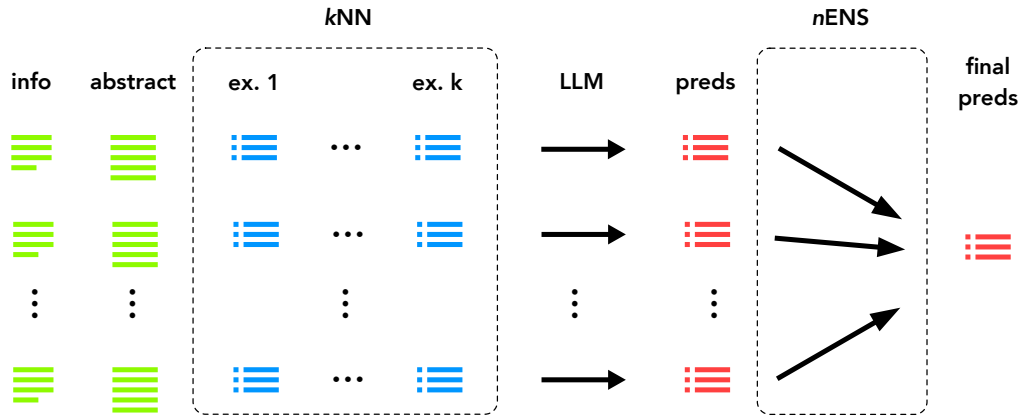

Figure 1: 2-step ICL approach: a kNN-based example prediction ($k = 3, 5$) step followed by an n-Ensembling ($n = 3$) step (cf. text for further details). For each abstract A, the class predictions $\hat{y}_1, dots, \hat{y}_m$ of all of its ACs $x_1, dots, x_m$ are generated in one inference step (all-at-once modality).

sample and generate the class label predictions for the ACs of an abstract all-at-once.

3.4. Implementation Details

As the embedding engine, we use dmis-lab's BioBERT[1]. For zero-shot learning, ICL and fine-tuning, we experiment with the LLaMA-3-8B-Instruct and LLaMA-3-70B-Instruct models, as well as various GGML-quantized configurations of them[2]. For ICL, we set the generate temperature to 0.1. For fine-tuning, we use LoRA adapters with `loraplus_lr_ratio` of 16.0. We set batch size of 2 and learning rate of $5e^{-5}$. For implementation, we use the LLaMA-Factory[3] framework [36]. An example of the prompts we use for zero-shot learning, in-context learning and fine-tuning with LLaMA-3 are given in Appendix A.

4. Results

4.1. Zero-Shot Learning

The results for zero-shot learning (ZSL) on ATC task are reported in Table 2. Recall that zero-shot learning corresponds to the prompting strategy where no nearest neighbors are included as demonstration examples, referred to as 'info + abstract + 0NN' in our notation. In an initial experimentation phase, we observed that adding complementary information (**info**) (definitions of 'Claim' and 'Premise' and dataset statistics) and including the entire text of the abstract (**abstract**) significantly improve the results. These expected observations serve as an ablation study and justify the usage of the additional information and full abstract text (prompt template 'info + abstract') in all subsequent experiments.

In all experiments, we observed that the models consistently generated the correct number of classes for each inference task. This observation remains valid for subsequent ICL and fine-tuning settings. It demonstrates the model's capability to understand the correspondence between the number of input ACs and the number of classes to predict.

In ZSL training-free setting, across Neo, Gla and Mix test sets, the performance of LLMs strongly correlated with the complexity of these models, achieving maximal macro F1-scores of 0.698, 0.819 and 0.725, respectively. Overall, in ZSL, the LLMs fail to achieve acceptable results. These considerations underscore the need for implementing additional learning modalities to address the ATC task effectively.

4.2. In-Context Learning

The results for in-context learning (ICL) on the ATC task are reported in Table 3. First, note that the transition from zero-shot learning ('info + abstract + 0NN', Table 2) to in-context learning ('info + abstract + kNN', Table 3) drastically improves the results. This validates the effectiveness of the kNNN-based examples selection method.

In addition, except for the Mix test set, the 3NN strategy consistently outperforms the 5NN strategy, suggesting that three examples suffice for optimal learning the ATC task in an ICL setting. The inclusion of more demonstration examples correlates with a significant increase

[1] https://huggingface.co/dmis-lab
[2] https://github.com/ggerganov/ggml
[3] https://github.com/hiyouga/LLaMA-Factory

Model	C	P	F1
Neo test			
LLaMA-3-8b-Instruct-bnb-4bit	0.529	0.539	0.534
LLaMA-3-8b-Instruct	0.544	0.558	0.551
LLaMA-3-70b-Instruct-bnb-4bit	0.642	0.753	**0.698**
Gla test			
LLaMA-3-8b-Instruct-bnb-4bit	0.553	0.635	0.594
LLaMA-3-8b-Instruct	0.569	0.692	0.631
LLaMA-3-70b-Instruct-bnb-4bit	0.755	0.882	**0.819**
Mix test			
LLaMA-3-8b-Instruct-bnb-4bit	0.546	0.524	0.535
LLaMA-3-8b-Instruct	0.563	0.564	0.563
LLaMA-3-70b-Instruct-bnb-4bit	0.671	0.779	**0.725**

Table 2
Zero-shot results for ATC on three test sets of the AbstRTC dataset using LLaMA-3.

in prompt length, potentially hindering the performance of the LLM or exceeding the maximum size of its context. Furthermore, the ensembling strategy consistently improves the results, even if only slightly, ensuring that the robustness of the results can indeed be strengthened through ensembling predictions.

Overall, the training-free ICL strategy achieves very competitive F1-scores of 0.912, 0.910, and 0.929 on Neo, Mix, and Gla test sets, respectively. However, these results remain lower than those obtained by previous training-dependent models (see Table 4, upper rows).

4.3. Fine-Tuning

The results achieved by the fine-tuning (FT) strategy on the ATC task are reported in Table 4. Our results show that fine-tuning significantly outperforms ICL. These findings suggest that the argumentative flow within abstracts cannot be inferred solely from the knowledge acquired during pre-training, and requires additional parameters updates to be effectively learned.

In this training-dependent context, we achieve maximal F1-scores of 0.935, 0.913, and 0.951 on the Neo, Gla, and Mix test sets, respectively, establishing new state-of-the-art results for the Neo and Mix test sets. These results suggest once again that the sequentiality of arguments inside a specific corpus requires fine-tuning to be optimally captured.

Prompt	C	P	F1
Neo test			
LLaMA-3-8b-Instruct			
info + abstract + 3NN	0.832	0.912	0.872
info + abstract + 5NN	0.843	0.914	0.878
info + abstract + 3NN + 3Ens	0.844	0.917	0.880
LLaMA-3-8b-Instruct-bnb-4bit			
info + abstract + 3NN	0.847	0.916	0.881
info + abstract + 5NN	0.817	0.890	0.853
info + abstract + 3NN + 3Ens	0.848	0.919	0.884
LLaMA-3-70b-Instruct-bnb-4bit			
info + abstract + 3NN	0.870	0.935	0.903
info + abstract + 5NN	0.863	0.930	0.896
info + abstract + 3NN + 3Ens	0.884	0.941	**0.912**
Gla test			
LLaMA-3-8b-Instruct			
info + abstract + 3NN	0.834	0.929	0.882
info + abstract + 5NN	0.836	0.925	0.881
info + abstract + 3NN + 3Ens	0.872	0.947	**0.910**
LLaMA-3-8b-Instruct-bnb-4bit			
info + abstract + 3NN	0.827	0.924	0.875
info + abstract + 5NN	0.816	0.916	0.866
info + abstract + 3NN + 3Ens	0.832	0.928	0.880
LLaMA-3-70b-Instruct-bnb-4bit			
info + abstract + 3NN	0.868	0.946	0.907
info + abstract + 5NN	0.865	0.945	0.905
info + abstract + 3NN + 3Ens	0.863	0.944	0.903
Mix test			
LLaMA-3-8b-Instruct			
info + abstract + 3NN	0.879	0.938	0.909
info + abstract + 5NN	0.898	0.944	0.921
info + abstract + 3NN + 3Ens	0.884	0.940	0.912
LLaMA-3-8b-Instruct-bnb-4bit			
info + abstract + 3NN	0.859	0.926	0.893
info + abstract + 5NN	0.866	0.922	0.894
info + abstract + 3NN + 3Ens	0.885	0.940	0.913
LLaMA-3-70b-Instruct-bnb-4bit			
info + abstract + 3NN	0.905	0.954	**0.929**
info + abstract + 5NN	0.906	0.952	0.929
info + abstract + 3NN + 3Ens	0.904	0.952	0.928

Table 3
Results for ATC on three test sets of AbstRCT dataset with LLaMA-3 models using the 2-step ICL strategy described in the text.

5. Conclusion

In this work, we address argument type classification (ATC) in the biomedical AbstRTC dataset with openly available LLaMA-3 from the three-fold perspective of zero-shot learning (ZSL), in-context learning (ICL) and fine-tuning (FT). We show that ZSL fails to achieve acceptable performance, ICL significantly improves the results, and FT reaches state-of-the-art performance.

These results support the fact that ATC task cannot be solved in a zero-shot setting by relying solely on general-purpose language modalities acquired during

Model	Neo	Gla	Mix
ResAttArg(Ensemble) [27]	0.879	0.877	0.897
SeqMT [28]	0.919	0.924	0.922
MRC_GEN [35]	0.928	0.926	0.940
GIAM [25]	0.930	**0.928**	0.936
LLaMA-3-8B-Instruct	0.919	0.908	0.939
LLaMA-3-8B-Instruct-bnb-4bit	**0.935**	0.910	0.953
LLaMA-3-70B-Instruct	0.929	0.913	0.940
LLaMA-3-70B-Instruct-bnb-4bit	0.921	0.908	**0.951**

Table 4
Fine-tuning results for ATC task on the three test sets of AbstRCT dataset using LLaMA-3.

pre-training. Additional learning is essential, either in the form of solved demonstration examples (ICL) or via parameters' updates (FT). We conjecture that the sequential flow of arguments within a text is a corpus-specific feature that cannot be inferred through zero-shot methods.

Previous works demonstrated that the text of argument components alone do not suffice to infer their argumentative roles [2, 4, 6]. Additional contextual, structural and syntactic features are necessary. In our ICL and FT settings, comprehensive contextual and structural information is incorporated through task-specific information and complete abstract text provided in the prompt. This information enables the model to discern the sequence of arguments, their associated markers, and other characteristics closely associated with their argumentative roles.

For future work, the design and implementation of a full AM pipeline using LLMs represents a major milestone. In this scenario, the LLM would take raw texts as input and produce a detailed map of the argumentative structure as output. We believe that LLMs will substantially transform the landscape of AM and its practical applications.

Acknowledgments

This work benefited from access to the computing resources of the L3i laboratory, operated and hosted by the University of La Rochelle. It is financed by the French government and the Region Nouvelle-Acquitaine. This research also benefited from institutional support RVO: 67985807 and partially supported by the grant of the Czech Science Foundation No. GA22-02067S. Finally, we are grateful to Playtika Ltd. for their support for this research.

References

[1] R. M. Palau, M.-F. Moens, Argumentation mining: The detection, classification and structure of arguments in text, in: Proceedings of ICAIL 2019, ICAIL '09, ACM, New York, NY, USA, 2009, pp. 98–107. URL: https://doi.org/10.1145/1568234.1568246. doi:10.1145/1568234.1568246.

[2] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Computational Linguistics 43 (2017) 619–659. URL: https://aclanthology.org/J17-3005. doi:10.1162/COLI_a_00295.

[3] P. Potash, A. Romanov, A. Rumshisky, Here's my point: Joint pointer architecture for argument mining, in: M. P. et al. (Ed.), Proceedings of EMNLP 2017, ACL, 2017, pp. 1364–1373. URL: https://doi.org/10.18653/v1/d17-1143. doi:10.18653/V1/D17-1143.

[4] T. Kuribayashi, H. Ouchi, N. Inoue, P. Reisert, T. Miyoshi, J. Suzuki, K. Inui, An empirical study of span representations in argumentation structure parsing, in: A. K. et al. (Ed.), Proceedings of ACL 2019, ACL, Florence, Italy, 2019, pp. 4691–4698. URL: https://aclanthology.org/P19-1464. doi:10.18653/v1/P19-1464.

[5] U. Mushtaq, J. Cabessa, Argument classification with BERT plus contextual, structural and syntactic features as text, in: M. T. et al. (Ed.), Proceedings of ICONIP 2022, volume 1791 of *CCIS*, Springer, 2022, pp. 622–633. URL: https://doi.org/10.1007/978-981-99-1639-9_52. doi:10.1007/978-981-99-1639-9_52.

[6] U. Mushtaq, J. Cabessa, Argument mining with modular BERT and transfer learning, in: Proceedings of IJCNN 2023, IEEE, 2023, pp. 1–8. URL: https://doi.org/10.1109/IJCNN54540.2023.10191968. doi:10.1109/IJCNN54540.2023.10191968.

[7] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. Wen, A survey of large language models, CoRR abs/2303.18223 (2023). URL: https://doi.org/10.48550/arXiv.2303.18223. doi:10.48550/ARXIV.2303.18223. arXiv:2303.18223.

[8] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, Z. Sui, A survey on in-context learning, CoRR abs/2301.00234 (2023). URL: https://doi.org/10.48550/arXiv.2301.00234. doi:10.48550/ARXIV.2301.00234. arXiv:2301.00234.

[9] H. Nori, et al., Can generalist foundation models outcompete special-purpose tuning? case study in medicine, CoRR abs/2311.16452 (2023). URL: https://doi.org/10.48550/arXiv.2311.16452. doi:10.48550/

ARXIV.2311.16452. arXiv:2311.16452.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. K. et al. (Ed.), Proceedings of NeurIPS 2022, volume 35, 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[11] S. Lei, G. Dong, X. Wang, K. Wang, S. Wang, Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework, CoRR abs/2309.11911 (2023). URL: https://doi.org/10.48550/arXiv.2309.11911. doi:10.48550/ARXIV.2309.11911. arXiv:2309.11911.

[12] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. arXiv:2203.11171.

[13] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on medical challenge problems, CoRR abs/2303.13375 (2023). URL: https://doi.org/10.48550/arXiv.2303.13375. doi:10.48550/ARXIV.2303.13375. arXiv:2303.13375.

[14] T. Mayer, Argument Mining on Clinical Trials, Theses, Université Côte d'Azur, 2020. URL: https://theses.hal.science/tel-03209489.

[15] R. Mochales, M. Moens, Argumentation mining, Artificial Intelligence and Law 19 (2011) 1–22. doi:10.1007/s10506-010-9104-x.

[16] I. Habernal, I. Gurevych, Argumentation mining in user-generated web discourse, Computational Linguistics 43 (2017) 125–179. URL: https://aclanthology.org/J17-1004. doi:10.1162/COLI_a_00276.

[17] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, N. Slonim, Context dependent claim detection, in: ICCL, 2014. URL: https://api.semanticscholar.org/CorpusID:18847466.

[18] S. Eger, J. Daxenberger, I. Gurevych, Neural end-to-end learning for computational argumentation mining, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of ACL 2017, ACL, Vancouver, Canada, 2017, pp. 11–22. URL: https://aclanthology.org/P17-1002. doi:10.18653/v1/P17-1002.

[19] V. Niculae, J. Park, C. Cardie, Argument mining with structured SVMs and RNNs, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of ACL 2017, ACL, Vancouver, Canada, 2017, pp. 985–995. URL: https://aclanthology.org/P17-1091. doi:10.18653/v1/P17-1091.

[20] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. B. et al. (Ed.), Proceedings of NAACL-HLT 2019, ACL, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/V1/N19-1423.

[21] G. Zhang, P. Nulty, D. Lillis, Enhancing legal argument mining with domain pre-training and neural networks, CoRR abs/2202.13457 (2022). URL: https://arxiv.org/abs/2202.13457. arXiv:2202.13457.

[22] H. Wang, Z. Huang, Y. Dou, Y. Hong, Argumentation mining on essays at multi scales, in: D. S. et al. (Ed.), Proceedings of COLING 2020, ICCL, Barcelona, Spain (Online), 2020, pp. 5480–5493. URL: https://aclanthology.org/2020.coling-main.478. doi:10.18653/v1/2020.coling-main.478.

[23] S. Fioravanti, A. Zugarini, F. Giannini, L. Rigutini, M. Maggini, M. Diligenti, Linguistic feature injection for efficient natural language processing, in: IJCNN 2023, June 18-23, 2023, IEEE, 2023, pp. 1–7. URL: https://doi.org/10.1109/IJCNN54540.2023.10191680. doi:10.1109/IJCNN54540.2023.10191680.

[24] J. Bao, C. Fan, J. Wu, Y. Dang, J. Du, R. Xu, A neural transition-based model for argumentation mining, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), ACL, Online, 2021, pp. 6354–6364. URL: https://aclanthology.org/2021.acl-long.497. doi:10.18653/v1/2021.acl-long.497.

[25] B. Liu, V. Schlegel, P. Thompson, R. T. Batista-Navarro, S. Ananiadou, Global information-aware argument mining based on a top-down multi-turn qa model, Information Processing & Management 60 (2023) 103445. URL: https://www.sciencedirect.com/science/article/pii/S0306457323001826. doi:https://doi.org/10.1016/j.ipm.2023.103445.

[26] A. Galassi, M. Lippi, P. Torroni, Argumentative link prediction using residual networks and multi-objective learning, in: N. Slonim, R. Aharonov (Eds.), Proceedings of the 5th Workshop on Argument Mining, ACL, Brussels, Belgium, 2018, pp. 1–10. URL: https://aclanthology.org/W18-5201. doi:10.18653/v1/W18-5201.

[27] A. Galassi, M. Lippi, P. Torroni, Multi-task attentive residual networks for argument mining, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 1877–1892. doi:10.1109/TASLP.2023.3275040.

[28] J. Si, L. Sun, D. Zhou, J. Ren, L. Li, Biomedical argument mining based on sequential multi-task learning, IEEE/ACM Trans. Comput. Biol. Bioinformatics 20 (2022) 864–874. URL: https://doi.org/

10.1109/TCBB.2022.3173447. doi:10.1109/TCBB.2022.3173447.

[29] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: G. D. G. et al. (Ed.), Proceedings of ECAI 2020, volume 325 of *FAIA*, IOS Press, 2020, pp. 2108–2115. URL: https://doi.org/10.3233/FAIA200334. doi:10.3233/FAIA200334.

[30] B. Molinet, S. Marro, E. Cabrio, S. Villata, T. Mayer, Acta 2.0: A modular architecture for multi-layer argumentative analysis of clinical trials, in: L. D. Raedt (Ed.), Proceedings of IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5940–5943. URL: https://doi.org/10.24963/ijcai.2022/859. doi:10.24963/ijcai.2022/859, demo Track.

[31] T. Mayer, S. Marro, E. Cabrio, S. Villata, Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials, Artificial Intelligence in Medicine 118 (2021) 102098. URL: https://www.sciencedirect.com/science/article/pii/S0933365721000919. doi:https://doi.org/10.1016/j.artmed.2021.102098.

[32] M. van der Meer, M. Reuver, U. Khurana, L. Krause, S. B. Santamaría, Will it blend? mixing training paradigms & prompting for argument quality prediction, in: G. Lapesa, et al. (Eds.), ArgMining@COLING 2022, ICCL, 2022, pp. 95–103. URL: https://aclanthology.org/2022.argmining-1.8.

[33] M. Pojoni, L. Dumani, R. Schenkel, Argument-mining from podcasts using chatgpt, in: L. Malburg, D. Verma (Eds.), Proceedings of ICCBR-WS 2023, volume 3438 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 129–144. URL: https://ceur-ws.org/Vol-3438/paper_10.pdf.

[34] A. Al Zubaer, M. Granitzer, J. Mitrović, Performance analysis of large language models in the domain of legal argument mining, Frontiers in Artificial Intelligence 6 (2023). URL: https://www.frontiersin.org/articles/10.3389/frai.2023.1278796. doi:10.3389/frai.2023.1278796.

[35] B. Liu, V. Schlegel, R. Batista-Navarro, S. Ananiadou, Argument mining as a multi-hop generative machine reading comprehension task, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. URL: https://openreview.net/forum?id=KTFxOnrbvu.

[36] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, Y. Ma, Llamafactory: Unified efficient fine-tuning of 100+ language models, in: Proceedings of the 62nd Annual Meeting of the ACL (Volume 3: System Demonstrations), ACL, Bangkok, Thailand, 2024. URL: http://arxiv.org/abs/2403.13372.

A. Appendix

Examples of prompts for LLaMA 3 for the zero-shot learning (ZSL), in-context learning (ICL) and fine-tuning (FT) settings are provided below.

A.1. Zero-Shot Learning

Task description: You are an expert biomedical assistant that takes 1) an abstract text and 2) the list of all arguments from this abstract text, and must classify all arguments into one of two classes: Claim or Premise. 68.0052% of examples are of type Premise and 31.9948% of type Claim. You must absolutely not generate any text or explanation other than the following JSON format {"Argument 1": <predicted class for Argument 1 (str)>, ..., "Argument n": <predicted class for Argument n (str)>}

Class definitions: Claim = A claim in the abstract of an RCT is a statement or conclusion about the findings of the study. Premise = A premise in the abstract of an RCT is a statement that provides an evidence or proof for a claim.

Abstract: Few controlled clinical trials exist to support oral combination therapy in pulmonary arterial hypertension (PAH). Patients with PAH (idiopathic [IPAH] or associated with connective tissue disease [APAH-CTD]) taking bosentan (62.5 or 125 mg twice daily at a stable dose for $\geq$3 months) were randomized (1:1) to sildenafil (20 mg, 3 times daily; n = 50) or placebo (n = 53). The primary endpoint was change from baseline in 6-min walk distance (6MWD) at week 12, assessed using analysis of covariance. Patients could continue in a 52-week extension study. An analysis of covariance main-effects model was used, which included categorical terms for treatment, baseline 6MWD (<325 m; $\geq$325 m), and baseline aetiology; sensitivity analyses were subsequently performed. In sildenafil versus placebo arms, week-12 6MWD increases were similar (least squares mean difference [sildenafil-placebo], -2.4 m [90% CI: -21.8 to 17.1 m]; P = 0.6); mean ± SD changes from baseline were 26.4 ± 45.7 versus 11.8 ± 57.4 m, respectively, in IPAH (65% of population) and -18.3 ± 82.0 versus 17.5 ± 59.1 m in APAH-CTD (35% of population). One-year survival was 96%; patients maintained modest 6MWD improvements. Changes in WHO functional class and Borg dyspnoea score and incidence of clinical worsening did not differ. Headache, diarrhoea, and flushing were more common with sildenafil. Sildenafil, in addition to stable ($\geq$3 months) bosentan therapy, had no benefit over placebo for 12-week change from baseline in 6MWD. The influence of PAH aetiology warrants future study.

Arguments: Argument 1=In sildenafil versus placebo arms, week-12 6MWD increases were similar (least squares mean difference [sildenafil-placebo], -2.4 m [90% CI: -21.8 to 17.1 m]; P = 0.6); mean ± SD changes from baseline were 26.4 ± 45.7 versus 11.8 ± 57.4 m, respectively, in IPAH (65% of population) and -18.3 ± 82.0 versus 17.5 ± 59.1 m in APAH-CTD (35% of population).
Argument 2=Changes in WHO functional class and Borg dyspnoea score and incidence of clinical worsening did not differ.
Argument 3=Headache, diarrhoea, and flushing were more common with sildenafil.
Argument 4=Sildenafil, in addition to stable ($\geq$3 months) bosentan therapy, had no benefit over placebo for 12-week change from baseline in 6MWD.

Result:

A.2. In-Context Learning (ICL)

Task description: You are an expert biomedical assistant that takes 1) an abstract text, 2) the list of all arguments from this abstract text, and must classify all arguments into one of two classes: Claim or Premise. 68.0052% of examples are of type Premise and 31.9948% of type Claim. You must absolutely not generate any text or explanation other than the following JSON format {"Argument 1": <predicted class for Argument 1 (str)>, ..., "Argument n": <predicted class for Argument n (str)>}

Class definitions: Claim = A claim in the abstract of an RCT is a statement or conclusion about the findings of the study. Premise = A premise in the abstract of an RCT is a statement that provides an evidence or proof for a claim.

Examples:

Example 1

Abstract:

Treatment of patients with advanced or metastatic esophagogastric adenocarcinoma should not only prolong life but also provide relief of symptoms and improve quality of life (QOL). Esophagogastric adenocarcinoma mainly occurs in elderly patients, but they are underrepresented in most clinical trials and often do not receive effective combination chemotherapy, most probably for fear of intolerance. Using validated instruments, we prospectively assessed QOL within the randomized FLOT65+ phase II trial. Within the FLOT65+ trial, a total of 143 patients aged $\geq$65 years were randomly allocated to receive biweekly oxaliplatin plus 5-fluorouracil (5-FU) continuous infusion and folinic acid (FLO) or the same regimen in combination with docetaxel 50 mg/m(2) (FLOT). The European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire C30 (EORTC QLQ-C30) and the gastric module STO22 were administered every 8 weeks until progression. Time to definitive deterioration of QOL parameters was analyzed and compared within the treatment arms. The median age of patients was 70 years. Patients receiving FLOT exhibited higher response rates and had improved disease-free and progression-free survival (PFS). The proportions of patients with evaluable baseline EORTC QLQ-C30 and STO22 questionnaires were balanced (83 % in FLOT and 89 % in FLO). Considering evaluable patients with assessable questionnaires (n = 123), neither functioning nor symptom parameters differed significantly in favor of one of the two treatment groups. Particularly, there was no significant difference regarding time to definitive deterioration of global health status/quality of life from baseline (primary endpoint). Notably, patients receiving FLO or FLOT as palliative treatment (n = 98) achieved comparable QOL results. Although toxicity was higher in patients receiving FLOT, no negative impact of the addition of docetaxel on QOL parameters could be demonstrated. Thus, elderly patients in need of intensified chemotherapy may receive FLOT without compromising patient-reported outcome parameters.

Arguments:

Argument 1=Patients receiving FLOT exhibited higher response rates and had improved disease-free and progression-free survival (PFS).
Argument 2=there was no significant difference regarding time to definitive deterioration of global health status/quality of life from baseline (primary endpoint).
Argument 3=patients receiving FLO or FLOT as palliative treatment (n = 98) achieved comparable QOL results.
Argument 4=Although toxicity was higher in patients receiving FLOT,
Argument 5=no negative impact of the addition of docetaxel on QOL parameters could be demonstrated.
Argument 6=elderly patients in need of intensified chemotherapy may receive FLOT without compromising patient-reported outcome parameters.

Result:

{"Argument 1": "Premise", "Argument 2": "Premise", "Argument 3": "Premise", "Argument 4": "Premise", "Argument 5": "Premise", "Argument 6": "Claim"}

Example 2

Abstract:

Chemotherapy prolongs survival and improves quality of life (QOL) for good performance status (PS) patients with advanced non-small cell lung cancer (NSCLC). Targeted therapies may improve chemotherapy effectiveness without worsening toxicity. SGN-15 is an antibody-drug conjugate (ADC), consisting of a chimeric murine monoclonal antibody recognizing the Lewis Y (Le(y)) antigen, conjugated to doxorubicin. Le(y) is an attractive target since it is expressed by most NSCLC. SGN-15 was active against Le(y)-positive tumors in early phase clinical trials and was synergistic with docetaxel in preclinical experiments. This Phase II, open-label study was conducted to confirm the activity of SGN-15 plus docetaxel in previously treated NSCLC patients. Sixty-two patients with recurrent or metastatic NSCLC expressing Le(y), one or two prior chemotherapy regimens, and PS< or =2 were randomized 2:1 to receive SGN-15 200 mg/m2/week with docetaxel 35 mg/m2/week (Arm A) or docetaxel 35 mg/m2/week alone (Arm B) for 6 of 8 weeks. Intrapatient dose-escalation of SGN-15 to 350 mg/m2 was permitted in the second half of the study. Endpoints were survival, safety, efficacy, and quality of life. Forty patients on Arm A and 19 on Arm B received at least one treatment. Patients on Arms A and B had median survivals of 31.4 and 25.3 weeks, 12-month survivals of 29% and 24%, and 18-month survivals of 18% and 8%, respectively Toxicity was mild in both arms. QOL analyses favored Arm A. SGN-15 plus docetaxel is a well-tolerated and active second and third line treatment for NSCLC patients . Ongoing studies are exploring alternate schedules to maximize synergy between these agents.

Arguments:

Argument 1=Chemotherapy prolongs survival and improves quality of life (QOL) for good performance status (PS) patients with advanced non-small cell lung cancer (NSCLC).
Argument 2=Targeted therapies may improve chemotherapy effectiveness without worsening toxicity.
Argument 3=Le(y) is an attractive target since it is expressed by most NSCLC.
Argument 4=SGN-15 was active against Le(y)-positive tumors in early phase clinical trials and was synergistic with docetaxel in preclinical experiments.
Argument 5=Patients on Arms A and B had median survivals of 31.4 and 25.3 weeks, 12-month survivals of 29% and 24%, and 18-month survivals of 18% and 8%, respectively
Argument 6=Toxicity was mild in both arms.
Argument 7=QOL analyses favored Arm A.
Argument 8=SGN-15 plus docetaxel is a well-tolerated and active second and third

line treatment for NSCLC patients

Result:

{"Argument 1": "Claim", "Argument 2": "Claim", "Argument 3": "Claim", "Argument 4": "Premise", "Argument 5": "Premise", "Argument 6": "Premise", "Argument 7": "Premise", "Argument 8": "Claim"}

Example 3

Abstract:

The impact of treatment on health-related quality of life (HRQoL) is an important consideration in the adjuvant treatment of operable breast cancer. Here we report mature HRQoL outcomes from the ATAC trial, comparing anastrozole with tamoxifen as primary adjuvant therapy for postmenopausal women with localized breast cancer. Patients completed the Functional Assessment of Cancer Therapy-Breast (FACT-B) questionnaire plus endocrine subscale (ES) at baseline, 3 and 6 months, and every 6 months thereafter. Baseline characteristics in the HRQoL sub-protocol were well balanced between the anastrozole (n = 335) and tamoxifen (n = 347) groups in the primary analysis population. As with previously published results at 2 years, there was no statistically significant difference in the Trial Outcome Index of the FACT-B, the primary endpoint of the study, between treatments at 5 years. There were no statistically significant differences between treatment groups in ES total scores. Consistent with the 2-year analysis, there were differences between treatment groups in patient-reported side effects: diarrhea (anastrozole 3.1% vs. tamoxifen 1.3%), vaginal dryness (18.5% vs. 9.1%), diminished libido (34.0% vs. 26.1%), and dyspareunia (17.3% vs. 8.1%) were significantly more frequent with anastrozole compared to tamoxifen. Dizziness (3.1% vs. 5.4%) and vaginal discharge (1.2% vs. 5.2%) were significantly less frequent with anastrozole compared to tamoxifen. In this, the first report of HRQoL over 5 years of initial adjuvant therapy with an aromatase inhibitor, we conclude that anastrozole and tamoxifen had similar impacts on HRQoL, which was maintained or slightly improved during the treatment period for both groups.

Arguments:

Argument 1=The impact of treatment on health-related quality of life (HRQoL) is an important consideration in the adjuvant treatment of operable breast cancer.
Argument 2=As with previously published results at 2 years, there was no statistically significant difference in the Trial Outcome Index of the FACT-B, the primary endpoint of the study, between treatments at 5 years.
Argument 3=There were no statistically significant differences between treatment groups in ES total scores.
Argument 4=there were differences between treatment groups in patient-reported side effects:
Argument 5=diarrhea (anastrozole 3.1% vs. tamoxifen 1.3%), vaginal dryness (18.5% vs. 9.1%), diminished libido (34.0% vs. 26.1%), and dyspareunia (17.3% vs. 8.1%) were significantly more frequent with anastrozole compared to tamoxifen.
Argument 6=Dizziness (3.1% vs. 5.4%) and vaginal discharge (1.2% vs. 5.2%) were significantly less frequent with anastrozole compared to tamoxifen.
Argument 7=In this, the first report of HRQoL over 5 years of initial adjuvant therapy with an aromatase inhibitor, we conclude that anastrozole and tamoxifen had similar impacts on HRQoL, which was maintained or slightly improved during the treatment period for both groups.

Result:

{"Argument 1": "Claim", "Argument 2": "Premise", "Argument 3": "Premise", "Argument 4": "Claim", "Argument 5": "Premise", "Argument 6": "Premise", "Argument 7": "Claim"}

Abstract:

Few controlled clinical trials exist to support oral combination therapy in pulmonary arterial hypertension (PAH). Patients with PAH (idiopathic [IPAH] or associated with connective tissue disease [APAH-CTD]) taking bosentan (62.5 or 125 mg twice daily at a stable dose for ≥3 months) were randomized (1:1) to sildenafil (20 mg, 3 times daily; n = 50) or placebo (n = 53). The primary endpoint was change from baseline in 6-min walk distance (6MWD) at week 12, assessed using analysis of covariance. Patients could continue in a 52-week extension study. An analysis of covariance main-effects model was used, which included categorical terms for treatment, baseline 6MWD (<325 m; ≥325 m), and baseline aetiology; sensitivity analyses were subsequently performed. In sildenafil versus placebo arms, week-12 6MWD increases were similar (least squares mean difference [sildenafil-placebo], -2.4 m [90% CI: -21.8 to 17.1 m]; P = 0.6); mean ± SD changes from baseline were 26.4 ± 45.7 versus 11.8 ± 57.4 m, respectively, in IPAH (65% of population) and -18.3 ± 82.0 versus 17.5 ± 59.1 m in APAH-CTD (35% of population). One-year survival was 96%; patients maintained modest 6MWD improvements. Changes in WHO functional class and Borg dyspnoea score and incidence of clinical worsening did not differ. Headache, diarrhoea, and flushing were more common with sildenafil. Sildenafil, in addition to stable (≥3 months) bosentan therapy, had no benefit over placebo for 12-week change from baseline in 6MWD. The influence of PAH aetiology warrants future study.

Arguments:

Argument 1=In sildenafil versus placebo arms, week-12 6MWD increases were similar (least squares mean difference [sildenafil-placebo], -2.4 m [90% CI: -21.8 to 17.1 m]; P = 0.6); mean ± SD changes from baseline were 26.4 ± 45.7 versus 11.8 ± 57.4 m, respectively, in IPAH (65% of population) and -18.3 ± 82.0 versus 17.5 ± 59.1 m in APAH-CTD (35% of population).
Argument 2=Changes in WHO functional class and Borg dyspnoea score and incidence of clinical worsening did not differ.
Argument 3=Headache, diarrhoea, and flushing were more common with sildenafil.
Argument 4=Sildenafil, in addition to stable (≥3 months) bosentan therapy, had no benefit over placebo for 12-week change from baseline in 6MWD.

Result:

A.3. Fine-Tuning (FT)

You are an expert in medical analysis. You are given the abstract of a random controlled trial which contains numbered argument components enclosed by <AC></AC> tags. Your task is to classify each argument components in the essay as either "Claim" or "Premise". You must return a list of argument component types in following JSON format: "component_types": [component_type (str), component_type (str), ..., component_type (str)]

Here is the abstract text: An open, randomized study was performed to assess the effects of supportive pamidronate treatment on morbidity from bone metastases in breast cancer patients. Eighty-one pamidronate patients and 80 control patients were monitored for a median of 18 and 21 months, respectively, for events of skeletal morbidity and the radiologic course of metastatic bone disease. The oral pamidronate dose was 600 mg/d (high dose [HD]) during the earliest study years, then changed to 300 mg/d (low dose [LD]) because of gastrointestinal toxicity. Twenty-nine of 81 pamidronate (HD/LD) patients first received 600 mg/d and were then changed to 300 mg/d; 52 of 81 pamidronate LD patients received 300 mg/d throughout the study. Tumor treatment was unrestricted. An overall intent-to-treat analysis was performed.<AC> In the pamidronate group, the occurrence of hypercalcemia, severe bone pain, and symptomatic impending fractures decreased by 65%, 30%, and 50%, respectively; event-rates of systemic treatment and radiotherapy decreased by 35% (P < or = .02). </AC><AC> The event-free period (EFP), radiologic course of disease, and survival did not improve. </AC><AC> Subgroup analyses suggested a dose-dependent treatment effect. </AC><AC> Compared with their controls, in pamidronate HD/LD patients, events occurred 60% to 90% less frequently (P < or = .03) and the EFP was prolonged (P = .002). </AC><AC> In pamidronate LD patients, event-rates decreased by 15% to 45% (P < or = .04). </AC><AC> Gastrointestinal toxicity of pamidronate caused a 23% drop-out rate, </AC><AC> but other cancer-associated factors seemed to contribute to this toxicity. </AC><AC> Pamidronate treatment of breast cancer patients efficaciously reduced skeletal morbidity. </AC><AC> The effect appeared to be dose-dependent. </AC><AC> Further research on dose and mode of treatment is mandatory. </AC>

{"component_types": ["Premise", "Premise", "Claim", "Premise", "Premise", "Premise", "Claim", "Claim", "Claim", "Claim"]}