

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

J  r  mie Cabessa^{1,2} Hugo Hernault³ Umer Mushtaq⁴

¹David Lab, University of Versailles (UVSQ) – Paris Saclay, Versailles, France

²Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

³Playtika Ltd., Lausanne, Switzerland

⁴L3i, University of La Rochelle, 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 focus 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 k NN-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.

Introduction

This work focuses on AM in the biomedical AbstrCT dataset [2]. We address the ATC task using quantized and unquantized openly available LLaMA-3 LLMs (cf. leaderboard). We experiment with zero-shot learning, in-context learning, 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 an ICL strategy that combines k NN-based example selection with majority vote ensembling [3]. 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.

Dataset

We consider the AbstrCT dataset which consists of abstracts of 650 Randomized Controlled Trials selected from the biomedical database PubMed.

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

Table 1. AbstrCT dataset statistics.

<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>

Methodology

Zero-Shot Learning (ZSL)

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.

In-Context Learning (ICL)

In-context learning (ICL) refers to ability of LLMs to learn how to solve a task based on a few example solutions provided in the prompt. We introduce a 2-step ICL strategy for argument type classification (ATC).

1. k NN-based examples selection ($k = 3, 5$): (i) $2k$ neighboring abstracts A_1, \dots, A_{2k} of A are selected according to the BioBERT embedding cosine similarity measure. (ii) k abstracts, A_{i_1}, \dots, A_{i_k} , are randomly chosen from A_1, \dots, A_{2k} . (iii) A prompt containing all the ACs and their corresponding classes in these k abstracts is constructed (k NN). (iv) The LLM predicts the classes $\hat{y}_1, \dots, \hat{y}_m$ of c_1, \dots, c_m on the basis of on this prompt.
2. n -Ensembling ($n = 3, 5$): (i) The k NN-based examples selection step, which involves randomness, is repeated n times (n Ens), leading to a set of n sequences of class predictions $\{(\hat{y}_{i,1}, \dots, \hat{y}_{i,m}) : i = 1, \dots, n\}$. (ii) The final class predictions $\hat{y}_1, \dots, \hat{y}_m$ of c_1, \dots, c_m are obtained by applying a component-wise majority vote to the n predictions sequences.

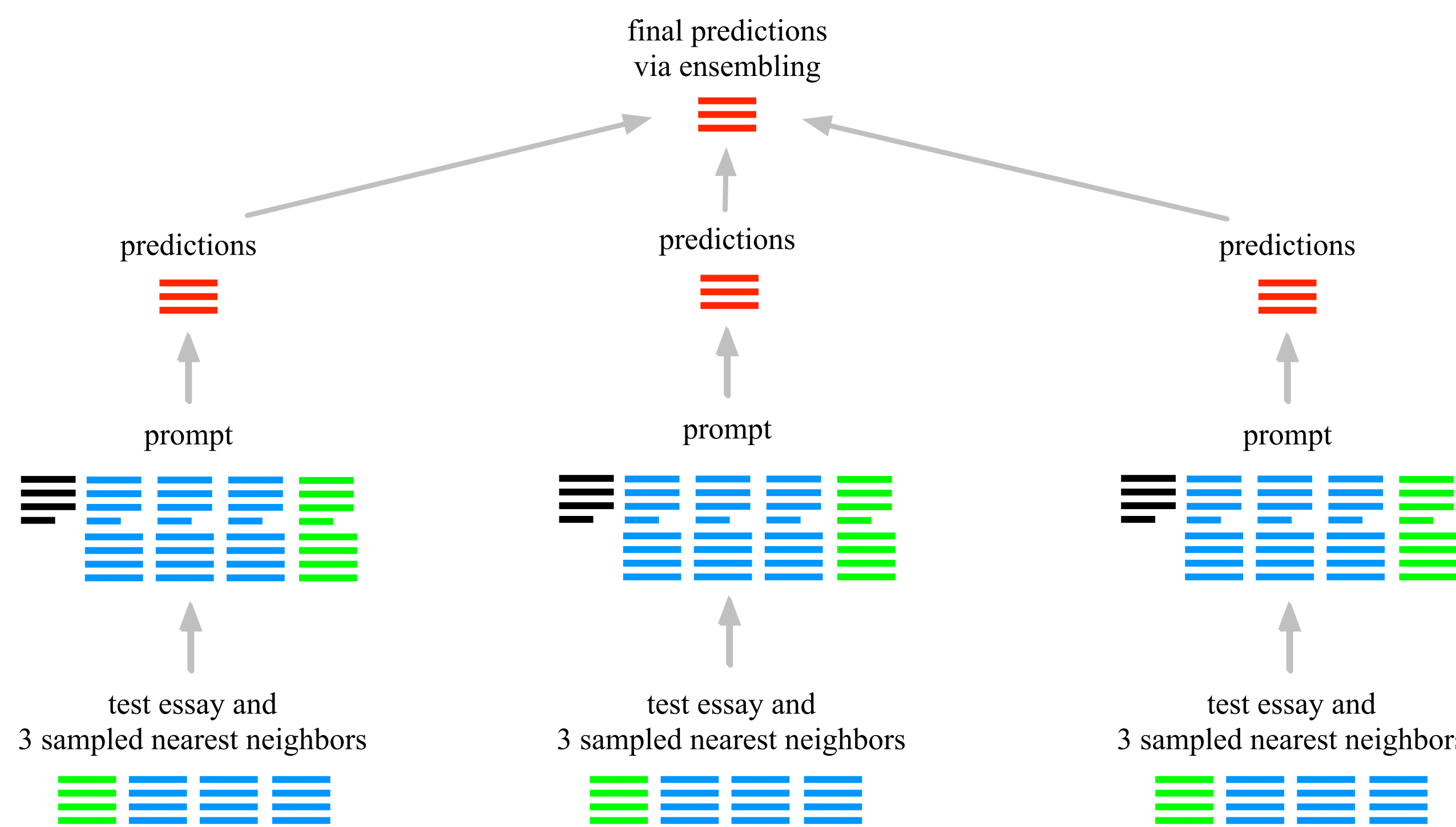


Figure 1. 2-step ICL approach: a k NN-based example prediction step ($k = 3$) followed by an n -Ensembling step ($n = 3$). For each abstract A , the class predictions of the of its ACs are generated all-at-once.

Fine-Tuning (FT)

Fine-tuning (FT) refers to the process of further training a pre-trained LLM on a downstream task. We propose a fine-tuning strategy that models the ATC task at the document level (all ACs are predicted at once).

Conclusion

- We addressed argument type classification (ATC) in the AbstrCT dataset with openly available LLaMA-3 models and using three approaches: zero-shot learning (ZSL), 2-step in-context learning (ICL) (new) and fine-tuning (FT).
- ZSL fails to achieve acceptable performance, ICL significantly improves the results, and FT reaches state-of-the-art performance.

Results

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. ZSL results for the ATC task.

Prompt	C	P	F1
Neo test			
LLaMA-3-8b-Instruct			
info + abstract + 3NN	0.832	0.912	0.872
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 + 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 + 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 + 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 + 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 + 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 + 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 + 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 + 3NN + 3Ens	0.904	0.952	0.928

Table 4. 2-step ICL results for the ATC task.

References

- [1] Boyang Liu, Viktor Schlegel, Paul Thompson, Riza Theresa Batista-Navarro, and Sophia Ananiadou. Global information-aware argument mining based on a top-down multi-turn qa model. *Information Processing & Management*, 60(5):103445, 2023.
- [2] Tobias Mayer. *Argument Mining on Clinical Trials*. Theses, Universit   C  te d’Azur, December 2020.
- [3] H. et al. Nori. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *CoRR*, abs/2311.16452, 2023.