# MAMET at SemEval-2024 Task 7: Supervised Enhanced Reasoning Agent Model

**Mahmood Kalantari**
Iran University of Science and Technology (IUST)
m_kalantari76@comp.iust.ac.ir

**Mehdi Feghhi**
Iran University of Science and Technology (IUST)
feghhi_me@comp.iust.ac.ir

**Taha Khany Alamooti**
Iran University of Science and Technology (IUST)
khany_taha@comp.iust.ac.ir

## Abstract

In the intersection of language understanding and numerical reasoning, a formidable challenge arises in natural language processing (NLP). Our study delves into the realm of NumEval, focusing on numeral-aware language understanding and generation using the QP, QQA and QNLI datasets[1]. We harness the potential of the Orca2 model, Fine-tuning it in both normal and Chain-of-Thought modes with prompt tuning to enhance accuracy. Despite initial conjectures, our findings reveal intriguing disparities in model performance. While standard training methodologies yield commendable accuracy rates. The core contribution of this work lies in its elucidation of the intricate interplay between dataset sequencing and model performance. We expected to achieve a general model with the Fine Tuning model on the QP and QNLI datasets respectively, which has good accuracy in all three datasets. However, this goal was not achieved, and in order to achieve this goal, we introduce our structure 1.

## 1 Introduction

In the realm of natural language understanding (NLU), the quest for models capable of comprehending and reasoning with textual data has been a longstanding pursuit. The NumEval task, focusing on Numeral-Aware Language Understanding and Generation, stands at the frontier of this endeavor, challenging researchers to develop models adept at grasping numerical information embedded within linguistic contexts. In this study, we delve into the intricacies of fine-tuning methodologies and their impact on the performance of language models, particularly focusing on the QP, QQA and QNLI datasets. (num)

The primary challenge in NLU lies in imbuing models with the ability to interpret and reason with textual information akin to human cognition. Traditional approaches often face hurdles in capturing the nuances of language, especially when numerical data intertwines with linguistic expressions. One possible cause of this problem is that numerals can have various notations, some of which are difficult to understand from their subwords. While models like Orca2 (Mitra et al., 2023), an instance of Large Language Models (LLMs), exhibit remarkable capabilities, their performance nuances in understanding numeral-aware contexts warrant deeper exploration.

The QP, QQA and QNLI datasets(Chen et al., 2023a), (Chen et al., 2019), (Ravichander et al., 2019), (Mishra et al., 2022) serve as test for evaluating the efficacy of language models in understanding questions, question-answering and natural language inference, respectively. These datasets present a diverse array of linguistic challenges, including numeral-aware reasoning, prompting the need for sophisticated training strategies.

Our study uses the Orca2 model, an advanced LLM known for its language comprehension skills. Through meticulous fine-tuning and evaluation on the QP, QQA and QNLI datasets.

We aim to catalyze discourse and innovation in the field of NLU, steering towards more robust and nuanced language models capable of navigating the complexities of numeral-aware language understanding and generation in real-world scenarios.

However, the road to achieving robust language understanding is fraught with challenges, chief among them being the inherent ambiguity and variability present in natural language. Numerical information adds an additional layer of complexity, requiring models to not only parse linguistic constructs but also interpret and reason with numerical data embedded within textual contexts.

Conventional training methodologies, while effective to a certain extent, often fall short in encapsulating the intricate interplay between linguistic semantics and numerical reasoning. The advent of large-scale language models has undoubtedly propelled the field forward, but their performance on numeral-aware tasks remains an area ripe for exploration and refinement.

After conducting several experiments, we have determined that fine-tuning a model for a specific subtask yields significantly higher accuracy compared to fine-tuning a model across all subtasks. Our attempt to fine-tune a generalized model across all subtasks while maintaining accuracy proved unsuccessful. Upon reviewing the results, we recognized the effectiveness of the Orca 2 model utilizing the LORA method for each subtask. Consequently, we trained the model using QLORA, resulting in improved accuracy. To establish a robust framework for addressing a range of reasoning

[1] https://drive.google.com/drive/folders/1mKbiL420U4Ih-hGmpaSki0FCvGHH3Au2?usp=sharing

*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1058–1063
June 20-21, 2024 ©2024 Association for Computational Linguistics

subtasks, we propose a structured approach that employs an agent as a supervisor capable of categorizing subtasks. This agent determines which of our fine-tuned models should address each task.

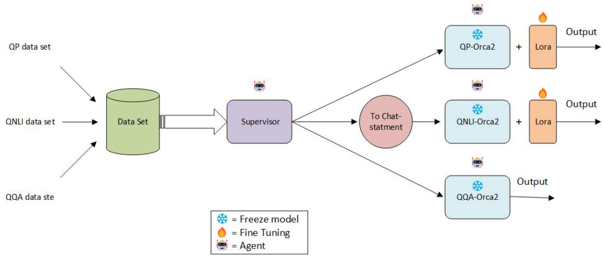

Figure 1: Supervised Enhanced Reasoning Agent Model.

## 2 Related Work

The intersection of Numerical Evaluation (NumEval) and Natural Language Processing (NLP) has witnessed a surge in research endeavors aimed at refining models' numerical comprehension and processing capabilities. Within this burgeoning field, a multitude of studies have delved into diverse methodologies and frameworks to deepen our understanding of numeracy in linguistic contexts.(Chen et al., 2023b)

A seminal study by (Chen et al., 2021) introduced the utilization of digit-based encoders to represent numerals, laying the groundwork for subsequent investigations into numerical representation methods. Expanding upon this, (Zhang et al., 2020) pioneered the exploration of scientific notation for numerical representation, shedding light on the efficacy of alternative numerical formats in quantitative skill tasks. These methodologies not only deepen our understanding of numerals' linguistic representation but also pave the way for novel approaches in numerical evaluation tasks.

Pretraining tasks have played a pivotal role in enhancing language models' capabilities in comprehending numerical language. (Devlin et al., 2019) revolutionized the field by introducing masked language model (MLM) and next sentence prediction (NSP) tasks, ushering in a new era of transformer-based NLP research. Building upon this foundation, (Yasunaga et al., 2022) proposed the document relation prediction (DRP) task, specifically designed to bolster models' performance in multi-hop reasoning and multi-document understanding tasks. These pretraining paradigms have significantly enriched models' numerical understanding capabilities, underscoring the pivotal role of pretraining in enhancing NLP models' numerical acumen.

In parallel, pre-finetuning strategies have emerged as a promising avenue for enhancing models' numerical comprehension. The Comparative Numbers Dataset (CND), introduced as a pre-finetuning resource, has garnered attention for its efficacy in enhancing models' numerical reasoning abilities. Experimental investigations leveraging BERT, RoBERTa, LinkBERT and FinBERT (Araci, 2019) have demonstrated notable improvements in models' performance across various numerical evaluation tasks, underscoring the potential of pre-finetuning methodologies in augmenting models' numerical understanding.(Chen et al., 2023b)

Generally, the landscape of NumEval research is characterized by a dynamic interplay of numerical representation methods, pretraining tasks and pre-finetuning strategies, each contributing to the advancement of language models' proficiency in numerical understanding and processing.(Chen et al., 2023b)

The results of our tests are very promising in this field and on the tested datasets, the accuracy is higher than the accuracy of the reference article.

## 3 Approach

Through our exploration with the Fine Tuning model, we discovered commendable accuracy in each sub-task individually. Upon scrutinizing the test outcomes, a noteworthy observation emerged: encoding numerical values within the text as statement-char significantly boosts accuracy in the QNLI task. However, a pertinent challenge persists: determining the appropriate agent for input assignment.

To address this challenge, we devised a framework. Initially, the input undergoes classification by an agent, ensuring its allocation to the most suitable model. In the case of the QNLI task, we preprocess words into statement-char format before directing them to the designated agent.

Furthermore, leveraging the Orca model's remarkable 100% accuracy in the QQA task, we opted to employ the base model as our agent. This strategic decision underscores our commitment to optimizing task performance and model efficacy.

### 3.1 Baseline Model:

The Orca2 model, a variant of the large language model (LLM), served as the cornerstone of our experiments. Built upon state-of-the-art architecture, Orca2 harnesses the power of deep neural networks to comprehend and generate human-like text responses. Leveraging its pre-trained weights, we fine-tuned Orca2 on the task-specific datasets to imbue it with numeral-aware capabilities.

### 3.2 Training Modes:

- Baseline Model Training: Initially, we evaluated the performance of the Orca2 model on each dataset section without any specific fine-tuning. This provided us with a baseline accuracy metric for comparison with subsequent experiments.

- Normal Fine Tuning: In this mode, we fine-tuned the Orca2 model on the respective datasets using conventional prompt tuning techniques. The model was trained to understand numeral-rich contexts and generated responses accordingly.

- Chain-of-Thoughts tuning Method: As an extension of traditional fine-tuning, we explored the Chain-of-Thoughts tuning method to train Orca2. This approach encourages the model to retain contextual information across sequential examples, enhancing its ability to grasp complex numeral-related nuances.

# 4 Experiments

## 4.1 Data

The four Used datasets (QP, QQA, QNLI [2] and AWPNLI [3]), are all related to natural language processing tasks and have been widely used in research and benchmarking for various NLP models, particularly those based on deep learning.

These datasets are often utilized to evaluate the performance of NLP models, particularly those designed for tasks like question answering, paraphrase detection and natural language inference. They provide standardized benchmarks for assessing the capabilities of different models and techniques in handling these tasks effectively.

## 4.2 Evaluation method

In evaluating the performance of the NumEval model across the QP, QQA and QNLI datasets, we employ a series of evaluation metrics tailored to the specific characteristics of each dataset and the different training modes applied to the Orca2 model.

### 4.2.1 Evaluation metrics:

Accuracy, F1-score and Recall serves as the primary evaluation metric across all experiments conducted on the QP, QQA, QNLI and AWPNLI datasets. It represents the proportion of correctly classified instances over the total number of instances in the datasets.

### 4.2.2 Experimental Modes:

Three experimental modes are considered in the evaluation:

- Basic Orca2 model without any fine-tuning.
- Orca2 model fine-tuned.
- Orca2 model fine-tuned using the Chain-of-Thought method.
- Sequential fine-tuning of Orca2 model using QP and QNLI datasets

Cross-dataset generalization is evaluated by training the Orca2 model on one dataset and subsequently fine-tuning it on another dataset to assess the model's ability to transfer knowledge across domains.

By employing these evaluation methods, we aim to comprehensively assess the effectiveness of the NumEval framework in numeral-aware language understanding and generation tasks across diverse datasets and training modes.

## 4.3 Experimental details

In our study, we conducted experiments utilizing the NumEval framework, focusing on the QP, QQA and QNLI datasets to assess the performance of the Orca2 model. Below, we outline the experimental details for each dataset and the various modes of training and testing conducted.

## 4.4 Results

In this section, we present the quantitative results obtained from our experiments across the QP, QNLI and QQA datasets. We compare our results against baselines established by various models, including BERT, CN-BERT, LinkBERT, CN-LinkBERT, RoBERTa and CN-RoBERTa, each evaluated on different data modes: original, Digit-based and ScientificNotation.

In table 1 presents the powers claimed in the article.

In the following, we present the quantitative results of our experiments conducted on the QP, QQA and QNLI datasets using the Orca2 model in various training modes including normal Fine tuning and Chain-of-Thoughts tuning in table 2. Additionally, we discuss the implications of these results in relation to our initial hypotheses and the effectiveness of our approach.

In a series of experiments, the model was Fine Tuned and tested on QNLI dataset which has scientific numbers or numbers that the decimal part is removed by multiplying by a large number with multiples of 10 units.

The quantitative results of our experiments reveal several noteworthy findings. Firstly, in the QP dataset experiments, we observed a substantial improvement in accuracy from the baseline when employing both Normal Fine-tuning and Chain-of-Thoughts tuning methods. Notably, the Chain-of-Thoughts tuning approach yielded the highest accuracy at 97.25%, demonstrating the effectiveness of sequential reasoning in improving model performance.

In contrast, the experiments conducted on the QNLI dataset showed similar trends, with both normal Fine Tuning and Chain-of-Thoughts tuning methods outperforming the baseline accuracy. The Chain-of-Thoughts tuning method again exhibited superior performance, underscoring its efficacy in capturing nuanced relationships within the data.

We developed a classifier model capable of discerning prompts based on their respective dataset classes: QP, QQA, and QNLI. This classifier model serves to categorize prompts and subsequently directs them to the corresponding model tailored to handle the specific prompt class. This streamlined approach obviates the necessity for segregating the datasets, enhancing overall

[2] https://drive.google.com/drive/folders/1mKbiL420U4Ih-hGmpaSki0FCvGHH3Au2?usp=sharing

[3] https://drive.google.com/file/d/10JNRN6iI5u9ZbEJPUEq4LAksBPW3vzGG/view?usp=sharing

| Model | Mode | QP_Comment | QP_Headline | QNLI | QQA |
|---|---|---|---|---|---|
| BERT | Original | 70.44% | 57.46% | 99.91% | 53.20% |
| | Digit-based | 65.38% | 54.74% | 99.11% | 53.75% |
| | ScientificNotation | 65.31% | 55.99% | 99.56% | 53.24% |
| CN-BERT | Digit-based | 69.93% | 54.84% | 99.42% | 52.53% |
| | ScientificNotation | 64.87% | 56.40% | 99.42% | 66.63% |
| LinkBERT | Original | 68.81% | 55.70% | 99.91% | 54.14% |
| | Digit-based | 63.76% | 55.41% | 99.73% | 53.44% |
| | ScientificNotation | 65.81% | 56.05% | 99.82% | 54.33% |
| CN-LinkBERT | Digit-based | 68.61% | 54.44% | 100% | 50.44% |
| | ScientificNotation | 63.48% | 53.15% | 99.73% | 52.11% |
| RoBERTa | Original | 60.46% | 58.03% | 98.93% | 51.96% |
| | Digit-based | 69.25% | 57.65% | 99.91% | 51.96% |
| | ScientificNotation | 64.32% | 55.49% | 100% | 53.67% |
| CN-RoBERTa | Original | 86.86% | 77.29% | 99.94% | 50.71% |
| | Digit-based | 64.25% | 55.92% | 99.73% | 50.88% |
| | ScientificNotation | 60.28% | 54.85% | 99.47% | 52.27% |

Table 1: Accuracy Results of article Models (Chen et al., 2023b)

| Experiment | Training Mode | QP_Comment | QP_Headline |
|---|---|---|---|
| Experiment 1 | Baseline | 68.48% | 80.12% |
| Experiment 2 | Normal Fine Tuning | 96.12% | 97.65% |
| Experiment 3 | Chain-of-Thoughts Tuning | 75.83% | 82.79% |
| Experiment 4 | FT on QNLI of FT on QP | 83.81% | 82.58% |

Table 2: Test results for the Orca2 model on the QP dataset

| Experiment | Training Mode | Accuracy |
|---|---|---|
| Experiment 1 | Baseline | 31.82% |
| Experiment 2 | Normal Fine Tuning 1 epoch | 98.34% |
| Experiment 3 | Normal Fine Tuning 2 epoch | 99.52% |
| Experiment 4 | Chain-of-Thoughts Tuning 1 epoch | 58.19% |
| Experiment 5 | Chain-of-Thoughts Tuning 2 epoch | 61.32% |
| Experiment 6 | Baseline Normal Fine Tuning on QP | 32.82% |
| Experiment 7 | Baseline Chain-of-Thoughts Tuning on QP | 33.23% |

Table 3: Test results for the Orca2 model on the QNLI dataset

| Model | Accuracy (%) | F1 score (%) | Recall (%) |
|---|---|---|---|
| Normal Fine Tuning 1 epoch | 98.34% | 98.51% | 98.34% |
| Normal Fine Tuning 2 epoch | 99.52% | 99.52% | 99.53% |
| Chain-of-Thoughts Tuning 1 epoch | 58.19% | 55.1% | 58.19% |
| Chain-of-Thoughts Tuning 2 epoch | 61.32% | 55.64% | 61.32% |

Table 4: F1 score and Recall for the Orca2 model on the Prompt Tuning by char-QNLI on QNLI dataset

| Experiment | Training Mode | Accuracy (%) |
|---|---|---|
| Experiment 1 | Normal Fine Tuning on statement-sci-10e 1 epoch | 33.74% |
| Experiment 2 | Normal Fine Tuning on statement-sci-10e 2 epoch | 53.99% |
| Experiment 3 | Normal Fine Tuning on char-QNLI 1 epoch | 96.87% |
| Experiment 4 | Normal Fine Tuning on char-QNLI 2 epoch | 99.65% |
| Experiment 5 | Normal Fine Tuning on char-QNLI 3 epoch | 99.79% |

Table 5: Test results for the Orca2 model on the Normal Fine Tuning by char-QNLI on QNLI dataset

| Model | Accuracy (%) | F1 score (%) | Recall (%) |
|---|---|---|---|
| Normal Fine Tuning on char-QNLI 1 epoch | 96.87% | 96.86% | 96.86% |
| Normal Fine Tuning on char-QNLI 2 epoch | 99.65% | 99.64% | 99.64% |
| Normal Fine Tuning on char-QNLI 3 epoch | 99.76% | 99.79% | 99.76% |

Table 6: F1 score and Recall for the Orca2 model on the Fine Tuning by char-QNLI on QNLI dataset

| Experiment | Training Mode | Accuracy (%) |
|---|---|---|
| Experiment | Baseline | 100% |

Table 7: Test Results for Orca2 in QQA Dataset

efficiency and coherence in the evaluation process.This structure is shown in Figure 1.

We employed a two-step fine-tuning process. Initially, the Orca2 model underwent fine-tuning on the QP dataset. Subsequently, we further fine-tuned the model using the QNLI dataset. The sequential fine-tuning approach allowed the model to adapt to the nuances of each dataset progressively.

We use the Orca2 model as an agent to assign each task to the specific agent. To achieve this task, we used a part of the available datasets to train our agent and achieved 99.4% accuracy in assigning tasks. As a result, we reached 96.13% accuracy on QP dataset, 100% accuracy on QQA and 98.85% accuracy on QNLI.

## 5 Analysis

Our analysis has illuminated the intricate nature of numeral-aware language tasks, emphasizing the imperative for ongoing scrutiny and enhancement of model architectures and training methods. The insights derived from our investigation notably elucidate the performance and adaptability of the Orca2 model within the NumEval task domain.

Upon reflection, we determined that optimizing the generality of the structure entails employing expert agents for individual sub-tasks, facilitated by a supervisory agent to assign tasks effectively. This strategic adjustment yielded heightened accuracy across all sub-tasks, marking a significant advancement in our approach.

## 6 Conclusion

In conclusion, our analysis highlights the intricate interplay between dataset characteristics, model architectures, and training methodologies in the NumEval task. While achieving significant advancements in numeral-aware language understanding and generation, our study underscores the importance of comprehensive evaluations encompassing both quantitative metrics and qualitative assessments to unravel the complexities of numerical reasoning in natural language understanding tasks. Moving forward, further research into adaptive learning strategies and nuanced dataset annotations promises to enrich our understanding and advancement in numeral-aware language processing tasks.

## References

Semeval-2024 task 7: Numeral-aware language understanding and generation.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Nquad: 70,000+ questions for machine comprehension of the numerals in text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2925–2929.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600K: Learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6307–6313, Florence, Italy. Association for Computational Linguistics.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2023a. Improving numeracy by input reframing and quantitative pre-finetuning task. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 69–77, Dubrovnik, Croatia. Association for Computational Linguistics.

Chung-Chi Chen et al. 2023b. Improving numeracy by input reframing and quantitative pre-finetuning task. *Findings of the Association for Computational Linguistics: EACL 2023*.

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

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

A Mitra, L Del Corro, S Mahajan, A Codas, C Simoes, S Agarwal, X Chen, and A Razdaibiedina. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

# A Example Appendix

### A.0.1 More about data set:

In each dataset there is a column called statement-char and statement-sci-10e. statement-char has spaced between the numbers inside the text and statement-sci-10e has written the numbers inside the text in scientific form (Chen et al., 2023b).

### A.0.2 Additional Experiments:

We explored the integration of human-reasoning into the AWPNLI dataset by generating human-reasoning for 190 samples using ChatGPT.

Contrary to expectations, the inclusion of human-reasoning did not yield the anticipated accuracy improvements, highlighting the complexity of the task and potential limitations of our approach.