# Noot Noot at SemEval-2024 Task 7: Numerical Reasoning and Headline Generation

**Sankalp Bahad**[1]
IIIT Hyderabad
sankalp.bahad@research.iiit.ac.in

**Yash Bhaskar**[1]
IIIT Hyderabad
yash.bhaskar@research.iiit.ac.in

**Parameswari Krishnamurthy**[2]
IIIT Hyderabad
param.krishna@iiit.ac.in

## Abstract

Natural language processing (NLP) models have achieved remarkable progress in recent years, particularly in tasks related to semantic analysis. However, many existing benchmarks primarily focus on lexical and syntactic understanding, often overlooking the importance of numerical reasoning abilities. In this paper, we argue for the necessity of incorporating numeral-awareness into NLP evaluations and propose two distinct tasks to assess this capability: Numerical Reasoning and Headline Generation. We present datasets curated for each task and evaluate various approaches using both automatic and human evaluation metrics. Our results demonstrate the diverse strategies employed by participating teams and highlight the promising performance of emerging models like Mixtral 8x7b instruct. We discuss the implications of our findings and suggest avenues for future research in advancing numeral-aware language understanding and generation.

## 1 Introduction

Natural language processing (NLP) models have achieved impressive performance on a wide range of semantic analysis tasks in recent years. However, the majority of benchmarks used to evaluate these models, including past SemEval shared tasks, have focused predominantly on lexical and syntactic understanding, with little emphasis on numerical reasoning abilities. In this paper, we argue that comprehending and reasoning with numerical values expressed in text is vital for robust language understanding, and should be an integral part of NLP evaluations going forward (num).

We demonstrate across several application scenarios that a lack of numeracy can undermine model performance and result in erroneous output. As an illustration, fine-grained sentiment analysis, as explored in SemEval-2017 Task 5, relies heavily on distinguishing subtle differences in sentiment intensity. Anticipating a 30% stock price increase implies a markedly more positive outlook than a 3% rise. Without accounting for the differential impact of these numbers, sentiment analysis models may fail to capture such nuances.

Similarly, in legal judgment prediction settings like SemEval-2023 Task 6, sentencing decisions can hinge on numerical quantities - stealing $100,000 typically incurs harsher penalties than stealing $10. Clinical inference use cases such as SemEval-2023 Task 7 also require sensitivity to numbers like blood pressure readings, where contrasts between 121 mmHg and 119 mmHg could indicate notably different health outlooks (Devlin et al., 2019).

These examples highlight the limitations of current benchmarking paradigms in evaluating true language comprehension. Accordingly, we propose that new numerically-grounded NLP tasks be developed to test numerical reasoning capacities. Recent work has begun exploring this direction, but substantial efforts are still needed to build robust models that demonstrate human-like numeracy. We outline a potential experimental framework and novel dataset for this purpose in the following sections.

## 2 Dataset

Task 3 (Huang et al., 2023) of our study comprises two distinct subtasks: numerical reasoning and headline generation. We describe the dataset for each subtask separately below:

### 2.1 Subtask 1: Numerical Reasoning

For the numerical reasoning subtask, we curated a dataset consisting of news headlines with missing numerical values. Each instance in the dataset includes a news article along with a headline where a numerical value is replaced with a blank. An example of the format for each instance is as follows:



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 913–917

- **News Article**: [Insert news article text here]
- **Headline with Blank**: "Study predicts a [blank] increase in global temperatures by 2050."
- **Target Value**: The correct numerical value that should fill in the blank.

The dataset includes a diverse range of news articles covering various topics such as climate change, economics, healthcare, and more. Each instance is associated with a target value representing the correct numerical answer.

### 2.2 Subtask 2: Headline Generation

For the headline generation subtask, we compiled a dataset consisting of news articles without headlines. Each instance in this dataset includes a news article, and the task is to generate a headline based on the content of the article. An example instance is provided below:

- **News Article**: [Insert news article text here]
- **Target Headline**: The headline that should be generated based on the content of the news article.

Similar to the numerical reasoning dataset, the articles cover a wide range of topics to ensure diversity in the generated headlines.

The table below shows the number of data points in the validation, test, and train sets for the Numerical Reasoning task:

| Dataset | Validation | Test | Train |
|---|---|---|---|
| Numerical Reasoning | 2572 | 4921 | 21157 |

Table 1: Dataset Statistics for Numerical Reasoning

Similarly, the table below presents the number of data points in the validation, test, and train sets for the Headline Generation task:

| Dataset | Validation | Test | Train |
|---|---|---|---|
| Headline Generation | 2365 | 5227 | 21157 |

Table 2: Dataset Statistics for Headline Generation

## 3 Methods

Zero-shot prompting is an effective method for these news headline tasks because it allows the model to apply its generalized language understanding capabilities to novel tasks without extensive fine-tuning. The model can deduce the appropriate responses based solely on the instructions and examples provided in the prompt.

The prompts are carefully engineered to provide the model with clear guidelines and context. For the numerical reasoning task, the prompt poses the incomplete headline as a question and asks the model to fill in the blank with only a numerical value. This focuses the model on extracting and inferring the relevant number from the article text.

Similarly, the headline generation prompt provides the news article as context and directly instructs the model to generate a headline summarizing the key information. The simplicity of these prompts allows the model to use its innate language skills to produce fitting responses without needing gradient-based training on the specific tasks.

Furthermore, the varied topics and contexts in the dataset require the model to adapt its numerical and summarization strategies across different situations. This tests the model's ability to generalize based on the prompt instructions, rather than overfitting to biases in a narrow dataset. The broad applicability demonstrated through zero-shot prompting highlights the versatile reasoning capacity gained through the model's pretraining.

Overall, zero-shot prompting is an elegant and effective approach for this study, as it allows assessment of the model's intrinsic skills at numerical deduction and text summarization when provided suitable prompts. The prompt formulation is key to eliciting successful performance without task-specific fine-tuning.

### 3.1 Subtask 1: Numerical Reasoning

The prompt (Zamfirescu-Pereira et al., 2023) used for determining the value of the missing numerical variable is as follows:

```
message = "News: {News}.\n
Headline: {Headline}\n\n
What is the value of ___?
Only give a numerical Response: "

prompt = f"[INST] {message} [/INST]"
```

### 3.2 Subtask 2: Headline Generation

For the headline generation subtask, participants are required to generate a headline based on the provided news article. The prompt for headline generation is as follows:

**Example 1: Numerical Reasoning**

**News:** (Oct 1, 2009 3:30 PM CDT) Want to catch up on YouTube2019s greatest hits but don2019t have the time? No problem: Just watch a new 4-minute mash-up that brings 100 of the best (or worst, depending on your viewpoint) together. Clips include such classics as Keyboard Cat and David After Dentist, and stars range from Obama Girl to the Dr. Pepper Guys, Time reports. Watch it at left.
**Masked Headline:** Watch 100 YouTube Classics in ____ Minutes
**Calculation:** Copy(4)
**Ans:** 4

**Example 2: Numerical Reasoning**

**News:** (Nov 16, 2009 8:40 AM) A rocket attack intended for a French general instead killed three children and wounded 20 others in a busy market northeast of Kabul today. Insurgents fired into the marketplace hoping to hit a meeting between Brig. Gen. Marcel Druart and tribal elders from Tagab Valley, where France is in the midst of a major offensive. Neither Druart nor any of his troops were harmed.
**Masked Headline:** Afghan Rocket Misses French General, Kills ____ Kids
**Calculation:** Trans(three)
**Ans:** 3

**Example 3: Numerical Reasoning**

**News:** (Mar 12, 2009 3:19 PM CDT) Stocks rose steadily after a morning dip today, with the Dow closing back over 7,000 points, MarketWatch reports. Bank of America and General Motors shot up 18.2% and 14.5%, respectively, after each announced they don2019t expect to ask the government for more bailout cash. The Dow ended up 239.66 at 7,170.06. The Nasdaq rose 54.46, settling at 1,426.10; the S
P 500 closed up 29.38 at 750.74.
**Masked Headline:** Dow Up 240, Retakes ____K Mark
**Calculation:** Paraphrase(7,000,K)
**Ans:** 7

**Example 4: Headline Generation**

**News:** (Mar 25, 2009 12:30 PM CDT) What's Italian for leadfoot ?00a0A Milanese man going 168mph was busted on four separate highway cameras in less than hour, ANSA reports. He was driving for his employer, whose lawyers argue that he should be responsible for just one infraction. They said they also plan to cite a court ruling that says signs identifying cameras must be a certain distance from a speed trap, adding: Naturally, we do not condone such driving at all.
**Headline:** Italian Going 168mph Gets 4 Tickets in 1 Hour

**Example 5: Headline Generation**

**News:** (Apr 21, 2010 12:51 PM CDT) The $100 bill is getting a new look and two high-tech security features to curb counterfeiters, the AP reports. A 3D security ribbon on the front has images of bells and 100s that move as you tilt the bill. The note, which is out next February, also has a Liberty Bell that seems to disappear. The government has more details here, along with a video that borders on the cheesy side here.
**Headline:** $100 Bill Goes 3D

```
message = "News: {News}.
Generate a headline based on the provided
news article: "

prompt = f"[INST] {message} [/INST]"
```

## 4 Results

### 4.1 Numerical Reasoning

The table 3 presents the results for the Numerical Reasoning task:

| Rank | Team | Score |
|---|---|---|
| 1 | CTYUN-AI | 0.95 |
| 2 | zhen qian | 0.94 |
| 3 | YNU-HPCC | 0.94 |
| 4 | NCL_NLP | 0.94 |
| 5 | NumDecoders | 0.91 |
| 6 | Infrrd.ai | 0.90 |
| 7 | hc | 0.88 |
| 8 | NLPFin | 0.86 |
| 9 | NP-Problem | 0.86 |
| 10 | AlRah | 0.83 |
| 11 | Noot Noot | 0.77 |
| 12 | GPT-3.5 | 0.74 |
| 13 | Sina Alinejad | 0.74 |
| 14 | StFX-NLP | 0.60 |

Table 3: Numerical Reasoning Results

In the domain of Numerical Reasoning, we showed a commendable performance, achieving a score of 0.77, marginally surpassing the baseline score attributed to the GPT-3.5 model, which stood at 0.74.

### 4.2 Headline Generation

#### 4.2.1 Auto Evaluation

The table 4 presents the results for the Headline Generation task based on auto evaluation metrics.

We demonstrated notable proficiency in headline generation, attaining a ROUGE score of 38.4, a BERT score of 57.5, and a Mover's Accuracy score of 3.6 in the automated evaluation.

#### 4.2.2 Human Evaluation

The table 5 presents the results for the Headline Generation task based on human evaluation metrics.

Human evaluators accorded the team a score of 1.68, indicating favorable reception of the generated headlines.

| Team | Num Acc. | ROUGE | BERT Score | Mover Score |
|---|---|---|---|---|
| ClusterCore | 38.2 | 51.6 | 13.9 | 33.5 |
| Noot Noot | 38.4 | 57.5 | 3.6 | 31.5 |
| Infrrd.ai | 65.8 | 68.4 | 61.3 | 46.8 |
| np_problem | 73.5 | 76.9 | 67.3 | 39.8 |
| hinoki | 62.4 | 66.3 | 55.2 | 43.1 |
| Challenges | 73.0 | 82.2 | 56.2 | 31.2 |
| NCL_NLP | 62.1 | 65.5 | 55.9 | 43.5 |
| YNU-HPCC | 69.0 | 73.0 | 61.8 | 48.9 |
| NoNameTeam | 55.7 | 57.7 | 52.1 | 40.7 |

Table 4: Auto Evaluation Results for Headline Generation

| Team | Num Acc. (50 Headlines) | Recommendation(100 News) |
|---|---|---|
| ClusterCore | 1.60 | 31 |
| Noot Noot | 1.68 | 11 |
| Infrrd.ai | 1.81 | 22 |
| np_problem | 1.57 | 14 |
| hinoki | 1.67 | 16 |
| Challenges | 1.70 | 10 |
| NCL_NLP | 1.73 | 16 |
| YNU-HPCC | 1.69 | 15 |
| NoNameTeam | 1.59 | 12 |

Table 5: Human Evaluation Results for Headline Generation

## 5 Conclusion

In this study, we explored the performance of various approaches for two distinct tasks: Numerical Reasoning and Headline Generation. Across both tasks, we observed a range of performances among the participating teams, indicating the diversity of techniques and strategies employed.

For Numerical Reasoning, the top-performing teams demonstrated high accuracy and effective reasoning capabilities, leveraging a combination of techniques to achieve superior results. Notably, the utilization of advanced models and fine-tuning methodologies played a crucial role in enhancing performance on this task.

In Headline Generation, both auto and human evaluations highlighted the effectiveness of certain teams in generating accurate and engaging headlines. Teams employing sophisticated natural language processing techniques, such as advanced neural models and feature engineering, exhibited superior performance in generating headlines that

resonated well with both automated evaluation metrics and human judges.

Furthermore, the introduction of Mixtral 8x7b instruct model showcased promising capabilities in both tasks. Despite not being fine-tuned specifically for the tasks at hand, the Mixtral model demonstrated competitive performance, particularly in Headline Generation. This suggests the robustness and versatility of the Mixtral 8x7b instruct model in understanding and generating natural language content across diverse domains and tasks.

Overall, our findings underscore the importance of leveraging state-of-the-art models and techniques in tackling complex natural language processing tasks. Additionally, the emergence of pretrained models like Mixtral 8x7b instruct offers promising avenues for future research and development, as they provide strong baselines and require minimal fine-tuning to achieve competitive performance across various NLP tasks.

## References


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

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.

Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Numhg: A dataset for number-focused headline generation. *arXiv preprint arXiv:2309.01455*.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.