SemEval-2024 Task 7: Numeral-Aware Language Understanding and Generation

Chung-Chi Chen
AIST, Japan
c.c.chen@acm.org

Jian-Tao Huang
Zhejiang Lab, China
jthuang@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Academia Sinica, Taiwan
hhhuang@iis.sinica.edu.tw

Hiroya Takamura
AIST, Japan
takamura.hiroya@aist.go.jp

Hsin-Hsi Chen
National Taiwan University, Taiwan
hhchen@ntu.edu.tw

Abstract

Numbers are frequently utilized in both our daily narratives and professional documents, such as clinical notes, scientific papers, financial documents, and legal court orders. The ability to understand and generate numbers is thus one of the essential aspects of evaluating large language models. In this vein, we propose a collection of datasets in SemEval-2024 Task 7 - NumEval. This collection encompasses several tasks focused on numeral-aware instances, including number prediction, natural language inference, question answering, reading comprehension, reasoning, and headline generation. This paper offers an overview of the dataset and presents the results of all subtasks in NumEval. Additionally, we contribute by summarizing participants' methods and conducting an error analysis. To the best of our knowledge, NumEval represents one of the early tasks that perform peer evaluation in SemEval's history. We will further share observations from this aspect and provide suggestions for future SemEval tasks.

1 Introduction

In the past, SemEval has predominantly focused on discussions surrounding words in text, with limited exploration of numbers in text. Recognizing the significance of understanding numbers can enhance performance in certain tasks. For instance, there is a notable difference in the sentiment degree between "expecting the stock price to increase by 30%" and "expecting the stock price to increase by 3%" in fine-grained sentiment analysis, as the former suggests a higher sentiment degree than the latter (SemEval-2017 Task 5 (Cortis et al., 2017)). Similarly, "Stealing \$10" versus "Stealing \$100,000" could result in differing court judgments (SemEval-2023 Task 6 (Modi et al., 2023)), and contrasting systolic blood pressure readings of 119 versus 121 offer different clinical inferences (SemEval-2023 Task 7 (Jullien et al., 2023)). These examples underscore the importance of numerical understanding in text, suggesting it as a potential research direction for enhancing the performance of downstream tasks.

Recent interest has surged in the numeracy of textual data and models within the NLP community, marking an opportune moment to evaluate current models' performance in numeral-aware language understanding and generation. To this end, we propose a collection of five published datasets encompassing three tasks: quantitative understanding, reading comprehension of numerals in text, and numeral-aware headline generation. For quantitative understanding tasks, we utilize the Quantitative 101 dataset (Chen et al., 2023). The NQuAD dataset (Chen et al., 2021) serves to explore reading comprehension with numerically rich documents, and Num-HG (Huang et al., 2024), annotated for numerical reasoning, facilitates the investigation of numeral-aware headline generation. In summary, while these are foundational NLP tasks, our focus is on discussing instances that require numeracy and the capacity to understand numbers for resolution.

In this paper, we first provide an overview of the dataset and subsequently summarize the methods and performances of participants. The comparison of models and error analysis will be included. Additionally, we employ peer evaluation to annotate and evaluate the generated outputs of participants' systems. Our analysis and observations, based on the annotations from participants, will be shared. We hope this pilot trial can offer insights and share experiences for future studies planning to conduct human evaluations among different teams.

2 Tasks and Datasets

We list the dataset for each task, the size and the corresponding license in Table 1. Quantitative 101, which is a collection of Numeracy-600K (Chen

Task	Subtask	Dataset	Size	Unit	License
Quantitative Understanding	Quantitative Prediction (QP)	Quantitative 101	1,200,000	Sentences	CC BY-NC-SA 4.0
	Quantitative Natural Language Inference (QNLI)		9,606	Sentence Pairs	MIT License
	Quantitative Question Answering (QQA)		807	Questions	ODC-By
Reading Comprehension of the Numerals in Text		NQuAD	71,998	News	CC BY-NC-SA 4.0
Numeral-Aware Headline Generation	Numerical Reasoning Headline Generation	Num-HG	27,746	News	CC BY-NC-SA 4.0

Table 1: Summary of the tasks and datasets in NumEval.

Subtask	Question	Answer
QP	FED'S DUDLEY REPEATS EXPECTS GDP GROWTH TO PICK UP IN 2014, FROM [Masked] PCT POST-RECESSION AVERAGE	1
QNLI	S1: Nifty traded above 7500, Trading Calls Today S2: Nifty above 7400	Entailment
QQA	Elliot weighs 180 pounds whereas Leon weighs 120 pounds. Who has a bigger gravity pull? Option1: Elliot Option2: Leon	Option 1

Table 2: Example for each subtask in Quantitative 101.

News Article: Major banks take the lead in self-discipline. The five major banks' newly-imposed mortgage interest rates climbed to **1.986%** in May. ... Also approaching **2%** integer alert ... Up to **2.5%** ... Also increased by **0.04** percentage points from the previous month ... Prevent the housing market bubble from fully starting.
Question Stem: Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly _____%.
Answer Options: (A) 0.04 (B) 1.986 (C) 2 (D) 2.5
Answer: (C)

Table 3: An example question in NQuAD.

News: At least **30** gunmen burst into a drug rehabilitation center in a Mexican border state capital and opened fire, killing **19** men and wounding **four** people, police said. Gunmen also killed **16** people in another drug-plagued northern city. The killings in Chihuahua city and in Ciudad Madero marked one of the bloodiest weeks ever in Mexico and came just weeks after authorities discovered **55** bodies in an abandoned silver mine, presumably victims of the country's drug violence. More than **60** people have died in mass shootings at rehab clinics in a little less than **two** years. Police have said **two** of Mexico's **six** major drug cartels are exploiting the centers to recruit hit men and drug smugglers, ...
Headline (Question): Mexico Gunmen Kill ____
Answer: 35
Annotation: Add(19,16)

Table 4: An annotation example in Num-HG.

et al., 2019), EQUATE (Ravichander et al., 2019), and NumGLUE Task 3 (Mishra et al., 2022). Some examples selected from these datasets are shown in Tables 2 and 3.[1] QP subtask aims to predict the magnitude of the masked number, and it is the coarse-grained setting for examing numeracy. QNLI and QQA subtasks require models to compare numbers to answer the question. RC task in NQuAD asks models to select a proper number for the question stem based on the given news article. The average of the micro-F1 score is used to evaluate the performance in Quantitative 101, and accuracy is used to evaluate the performance in NQuAD.

To go one step further, Num-HG extends the RC task in NQuAD. It provides numerical reasoning annotations to 27,746 news, and offers two subtasks, numerical reasoning and headline generation. The major goal of this task is to generate a headline that contains key numerical information in the news article. Table 4 shows an example of the Num-HG. In the numerical reasoning subtask, models need to calculate the correct number of the blank part in the news headline. In the headline generation subtask, models must generate a headline based on the given news. Because each headline in the proposed Num-HG contains one number, models are expected to generate the same number as journalists. Our rationale is that the number selected by the journalists should be the most informative for summarizing the news article. Therefore, we will evaluate whether the generated number is correct or not. Additionally, we will further evaluate the generated headline by automatical metrics, such as ROUGE and BERTScore, and manual evaluation. Specifically, participants manually evaluate the system outputs from other teams.

3 Participants and Automatic Evaluation

There are 124 teams registered for NumEval, with 20 teams submitting their system description papers. This section provides an overview of the major methods employed in each paper, with detailed explorations available in the respective papers. As participants can select specific subtasks, results are reported in a fine-grained manner to encompass partial outcomes.

Table 5 presents the Quantitative 101 results. Chen et al. (2024) utilize Flan-T5 (Chung et al., 2022) with an instructional prompt across all tasks,

[1]Examples in Tables 2, 3, and 4 are from the original papers.

Team	Method	QP		QNLI					QQA	Score
		comment	headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
YNU-HPCC	Flan-T5 + Instruction Prompt	67.20	58.82	77.73	52.40	77.06	68.40	99.94	59.25	70.10
HIJLI_JU	BERT + Character Representation	-	-	-	-	-	-	-	53.70	-
MAMET	Orca2	96.12	97.65			98.85			100.00	-
Calc-CMU	Pre-Calc (RoBERTa + Operation Classification + Calculator)	-	-	73.90	58.17	82.21	78.00	100.00	61.05	-
JU United	BERT	40.00		-	-	-	-	-	-	-
Bit_numeval	Abe-7B + Human Feedback	-	-	86.99	87.25	71.36	75.20	56.68	-	-

Table 5: Automatic Evaluation — Quantitative 101.

Team	Method	Accuracy
YNU-HPCC	Randeng-T5-77M	89.71
JN666	BERT + Pre-Finetuning with Companring Number Task	79.40
CYUT	BERT + Number Augumentation + Features	77.09

Table 6: Automatic Evaluation — NQuAD.

outperforming direct applications of pre-trained models such as BERT (Devlin et al., 2019), ReBERTa (Liu et al., 2019), and LinkBERT (Yasunaga et al., 2022). Sengupta et al. (2024) emphasize the significance of number representation in character format. Kalantari et al. (2024) employ Orca2 (Mitra et al., 2023) with fine-tuning and Chain-of-Thought (CoT) prompting (Wei et al., 2022), achieving high performance across most tasks. Veerendranath et al. (2024) introduce the Pre-Calc approach, which incorporates operation classification tasks during RoBERTa training and utilizes this knowledge to decide on calculator usage for results, highlighting the value of tool utilization. Saha (2024) experiment with BERT, while Liang et al. (2024) leverage the Abe-7B model enhanced by human feedback during training, surpassing several large language models (LLMs). In summary, findings from Quantitative 101 suggest that learning calculator usage and character-format number representation can aid in quantitative tasks. Additionally, employing tailored language models like Orca2 or integrating human feedback can further enhance performance.

Table 6 presents the results on the NQuAD dataset. Chen et al. (2024) achieved the highest performance using Randeng-T5-77M (Zhang et al., 2022). Liu et al. (2024), supporting previous research (Chen et al., 2023), demonstrated that pre-finetuning with a comparing numbers task could enhance performance. Lau and Wu (2024) introduced a numeral augmentation method to improve performance. In conclusion, a well-trained language model, such as Randeng-T5-77M, can achieve superior performance in reading comprehension tasks.

Table 7 presents the outcomes of the numerical reasoning task. Due to a few teams either missing the submission deadline or reporting their results in different formats, these are included in the unofficial evaluation section. For comprehensive details on their methodologies and results, their respective papers should be consulted. LLMs demonstrated commendable performance in this task, with the methodologies of the participants detailed subsequently. Fan et al. (2024) secured the highest performance with Qwen-72B-Chat (Bai et al., 2023), employing a strategy that distinguishes the input question as either a calculation or an application problem, alongside utilizing a data augmentation technique to enhance performance. Their approach incorporated two additional datasets: GSM8K (Cobbe et al., 2021) and MetaMathQA (Yu et al., 2023). Qian et al. (2024) disclosed the results of fine-tuning GPT-3.5, whereas Chen et al. (2024) applied Flan-T5 with Chain of Thought (CoT), complemented by the use of a calculator for accuracy improvement, which yielded superior results compared to direct arithmetic computations by models. Zhao et al. (2024) fine-tuned Mistral-7B (Jiang et al., 2023), achieving performance comparable to that of fine-tuned GPT-3.5. Gonzalez et al. (2024) combined the outputs of Flan T5 and GPT-3.5, whereas He et al. (2024) implemented Llama 2-7B (Touvron et al., 2023) with CoT. Additionally, Crum and Bethard (2024) utilized Flan-T5-Lamini, and Rajpoot and Chukamphaeng (2024) fine-tuned Mistral-7B. Bahad et al. (2024) reported the performance derived from prompting GPT-3.5. In conclusion, fine-tuning LLMs and a clear understanding of the task, particularly the decision on whether to employ an external calculator, are crucial for achieving enhanced performance in numerical reasoning tasks.

Table 8 displays the outcomes of headline generation tasks. Rajpoot and Chukamphaeng (2024) enhanced Mistral-7B, yielding headlines with numerals closely matching those chosen by journalists. In the reasoning subset, this approach also secures high accuracy. Chuang and Zhunis (2024) employed BART (Lewis et al., 2020) alongside a contractive learning approach, achieving superior performance in the copying subset. Compared to these

	Team	Method	Accuracy
Official	CTYUN-AI	Qwen-72B-Chat + Task Classification + Data Augmentation	0.95
	ZXQ	Finetuned GPT-3.5	0.94
	YNU-HPCC	Flan-T5 + CoT + Calculator	0.94
	NCL_NLP	Mistral-7B + CoT + Finetune	0.94
	NumDecoders	Ensemble (Flan T5 + GPT-3.5)	0.91
	Infrrd.ai	Llama 2-7B + CoT	0.90
	hc	Flan-T5-LaMini	0.88
	NP-Problem	Finetuned Mistral-7B	0.86
	AlRah	-	0.83
	Noot Noot	GPT-3.5	0.77
	Sina Alinejad	-	0.74
	StFX-NLP	-	0.60
Unofficial	VHA	DistilRoBERTa	-
	IUST-NLPLAB	GPT-3.5	-

Table 7: Automatic Evaluation — Numerical Reasoning.

	Team	Method	Num Accuracy			ROUGE			BERTScore			MoverScore
			Overall	Copy	Reasoning	1	2	L	P	R	F1	
Official	NP-Problem	Finetuned Mistral-7B	**73.49**	76.91	**67.26**	39.82	17.58	34.34	27.80	48.56	37.82	57.02
	Challenges	BART + Contrastive Learning	72.96	**82.17**	56.18	31.22	12.24	26.86	19.53	47.56	33.13	55.36
	YNU-HPCC	Flan-T5 + Instruction Tuning + Retrieved Similar Example	69.04	73.02	61.81	**48.85**	**24.68**	**44.18**	**51.55**	**50.10**	**50.38**	**60.55**
	Infrrd.ai	Llama 2-7B + RAG	65.84	68.35	61.26	46.79	22.36	42.10	51.01	47.26	49.13	59.73
	hinoki	T5-Based Title Generator	62.35	66.28	55.18	43.07	19.72	39.00	47.22	43.44	45.34	58.71
	NCL_NLP	Mistral-7B + CoT + Finetune	62.12	65.54	55.90	43.51	19.39	38.88	46.40	45.04	45.73	58.86
	NoNameTeam	-	55.72	57.68	52.13	40.65	17.26	35.75	44.26	40.39	42.32	57.74
	Noot Noot	GPT-3.5	38.39	57.48	3.63	31.47	11.14	27.28	25.39	43.98	34.54	55.56
	ClusterCore	Few-Shot Llama	38.23	51.57	13.94	33.47	11.84	28.93	31.88	42.23	37.03	56.41
Unofficial	VHA	T5	-	-	-	-	-	-	-	-	-	-

Table 8: Automatic Evaluation — Headline Generation.

teams, several groups have utilized LLMs, obtaining improved scores in ROUGE, BERTScore, and MoverScore metrics, albeit with reduced numeral precision. Chen et al. (2024) implemented Flan-T5 with an instruction tuning strategy and enhanced it by retrieving similar cases for model referencing, leading to top results across ROUGE, BERTScore, and MoverScore evaluations. He et al. (2024) applied Llama-2-7B with retrieval augmented generation (RAG), while Crum and Bethard (2024) developed a T5-based title generator.[2] Zhao et al. (2024) fine-tuned Mistral-7B using CoT, and Bahad et al. (2024) engaged GPT-3.5 through prompting. Singh et al. (2024) examined the efficacy of Llama under a few-shot learning framework. Overall, these findings suggest that while fine-tuning can enhance numeral selection accuracy, it might decrease the similarity between the generated headlines and the actual headlines.

4 Human Evaluation

4.1 Guidelines

To enhance the evaluation of generated headlines, we implement peer evaluation for the outputs from participants' systems. Participants are required to assess the models of other teams. The evaluation comprises two metrics:

- **Numerical Accuracy**: This metric evaluates the precision of numbers within the generated headlines. It aims to verify the correctness of numerical data presented in each headline. Systems are ranked based on their average scores, adhering to the following criteria:
  - Assign 2 points for fully accurate numerical data.
  - Allocate 1 point for partially accurate numbers.
  - Give 0 points for completely inaccurate or missing numbers.
- **Optimal Headline**: This assessment involves selecting the most appropriate headline from a set of nine options. Given that nine teams have submitted their outcomes for review, we substitute the outputs from the evaluating team with journalists' headlines, serving as the ground truth. The "best headline" is identified as the one that the evaluator considers most suitable for the journalist of the corresponding news article. The system receiving the highest number of votes will be awarded one point, with points accumulated for ranking purposes. If multiple systems tie with the

[2] https://huggingface.co/czearing/article-title-generator

Team	Numerical Accuracy	Optimal
Infrrd.ai	**1.81**	22
NCL_NLP	1.73	16
Challenges	1.70	10
YNU-HPCC	1.69	15
Noot Noot	1.68	11
hinoki	1.67	16
ClusterCore	1.60	**31**
NoNameTeam	1.59	12
NP_Problem	1.57	14
Ground Truth	-	28

Table 9: Human Evaluation

Operator	Ratio
Copy	23.42%
Trans	9.91%
Paraphrase	11.71%
Round	21.62%
Subtract	7.21%
Add	11.71%
Span	4.50%
Divide	4.50%
Multiply	5.41%

Table 10: Statistics of the operators present in the error sets of the top four systems for numerical reasoning.

same number of votes for first place on a given instance, each will receive one point.

4.2 Evaluation Results

Table 9 presents the outcomes of the human evaluation process. Numerical accuracy is derived from evaluating 50 instances, with each instance receiving three annotations. The determination of the optimal headline originates from the analysis of 100 instances. According to the results, He et al. (2024) secures the highest marks in terms of numerical accuracy, despite their fourth position in automatic evaluation. Furthermore, while Rajpoot and Chukamphaeng (2024) achieves the top rank in automatic evaluation, their performance is observed to be the least favorable in human assessment among all systems evaluated. An additional noteworthy observation is that Zhao et al. (2024), utilizing the same language model as Rajpoot and Chukamphaeng (2024), attains higher scores in human evaluation.

In the context of optimal headline generation, Singh et al. (2024) receives the highest score, even though it is placed at the lower end in automatic evaluation and does not exhibit exceptional performance in numerical accuracy. He et al. (2024) is ranked second in this regard, outperforming other teams. These findings suggest that Llama (2) excels in tasks related to headline generation, considering both numerical accuracy and optimal headline aspects. Given that the ground truth was also evaluated as a candidate, its score is disclosed in Table 9, where it achieves 28 points. This score is superior to most systems and marginally lower than that of Singh et al. (2024).

5 Discussion

5.1 Error Analysis

Through the examination of participant contributions, it is observed that simple numerical questions are on the verge of being effectively addressed with the selection of an optimal language model for specific tasks. In quantitative tasks, Kalantari et al. (2024) reports achieving over 96% in micro-F1 across all subtasks through the application of Qrca2. Within the NQuAD framework, Chen et al. (2024) employs Randeng-T5-77M to secure approximately 90% accuracy, while Fan et al. (2024) attains a 95% accuracy rate utilizing Qwen-72B-Chat. For the task of headline generation, numerous teams have recorded impressive scores in human evaluations, matching or surpassing the ground truth benchmarks. These findings suggest that the era may be approaching a point where traditional tasks requiring numerical understanding and generation are nearly resolved.

However, there remain several challenges for current language models. In Table 10, we provide statistics of the operators present in the error sets of the top four systems for numerical reasoning. For instance, when presented with the masked headline "Mother of 3 Gives Huge Gift to Dying Friend" based on the news:

> "When Beth Laitkep's breast cancer spread to her brain and spine, doctors realized she had limited time left. The concern arose about the future of her six children. 'If a miracle doesn't occur and I do not survive, could you take my children as your own?' she inquired of her friend Stephanie Culley, as recounted to People magazine. Culley agreed without hesitation. Consequently, Ace (aged 2), Lily (5), Dallas (10), Jaxson (11), Selena (14), and Will (15) moved in with Culley, her husband Donnie, and their three children following Laitkep's demise in May at 39. Fortuitously, Donnie, a construction worker, had constructed their home in Alton, Virginia, with ample bedrooms

	Infrrd.ai	NCL_NLP	Challenges	YNU-HPCC	Noot Noot	hinoki	ClusterCore	NoNameTeam	np_problem	Ground Truth
Infrrd.ai	-	11	9	9	**26**	3	15	3	12	12
NCL_NLP	19	-	0	13	0	23	0	20	1	**24**
Challenges	15	7	-	**22**	7	7	9	8	4	20
YNU-HPCC	**28**	15	1	-	0	18	5	12	5	16
Noot Noot	9	12	6	5	-	2	**31**	3	27	5
hinoki	8	9	11	5	**29**	-	23	4	6	5
ClusterCore	1	3	20	2	**70**	0	-	0	3	1
NoNameTeam	10	**18**	5	14	8	6	16	-	11	12
np_problem	8	8	14	15	6	7	12	10	-	**20**
Preferred	1	1	0	1	3	0	1	0	0	2

Table 11: Human Preference.

> to accommodate everyone. 'She is exceedingly humble and refrains from seeking assistance,' a friend of Stephanie's informed WSET. 'She's an angel.' (This family adopts children who are facing terminal conditions.)"

Three out of four models filled the blank with 6, while one model suggested 7. This instance illustrates the difficulty models face with numerical reasoning in complex narrative contexts.

Another intricate scenario involves a report that "A 66-year-old woman, pregnant and poised to become Britain's oldest mother, remains unrepentant about her choice, asserting her feeling akin to a 39-year-old on certain days," as detailed by the Mirror. Despite the varied daily feelings of being 39 or 56, Munro, who is 8 months pregnant, disregards the media attention, emphasizing the personal nature of her pregnancy decision. However, all models incorrectly predicted 39 instead of 66 for the headline "Brit Mum-to-Be 'Younger at Heart' Than 66, She Tells Critics".

Moreover, there are instances where models simply replicate rather than approximate numbers. For example, the correct answer for the headline "Car Auctions Off for Record-Breaking $____M" is 34.7, yet model predictions included 34.6, 34.65, and 38.0, with 34.65 being directly taken from the article text. In this case, some generated results may still be correct but just not the same as ground truth.

5.2 Human Preference

Given that most models, particularly LLMs, are adept at producing fluent headlines, the pertinent discussion revolves around the selection criteria among multiple headline candidates. This section delves into analyzing optimal headline annotations based on participant feedback. Table 11 presents statistics from different teams' annotations, highlighting the diversity in human preferences towards headline recommendations. Notably, most systems were primarily favored by a single team, with the exception of Bahad et al. (2024), which garnered the highest votes from three teams. Singh et al. (2024)'s pronounced preference for Bahad et al. (2024)'s system outputs stands out. Apart from this unique instance, determining the superior system is challenging, as preferences may vary across users. Another key observation is the ground truth achieving scores comparable to those of headlines generated by various systems, suggesting that striving for verbatim replication of the ground truth may be becoming obsolete in the context of LLMs. The emphasis may shift towards assessing the quality of generated text through more subjective and nuanced measures. Furthermore, the human evaluation results depicted in Table 11 underscore the difficulty in appraising generated headlines through manual voting, given the variance in team preferences. This inquiry constitutes the inaugural research question posited by NumEval, paving the way for subsequent investigations aimed at enhancing headline generation methodologies.

Aspect	Statistics
Average Length of Best Headlines	9.47 Words
Average Length of Other Selected Headlines	9.54 Words
ROUGE 1 between Best and Other	0.4373
ROUGE 2 between Best and Other	0.1951
ROUGE L between Best and Other	0.3791

Table 12: Statistics of the best headline (Best) and other selected headlines (Other).

To further elucidate, we present statistics in Table 12, computed based on headlines chosen by at least one annotator. Initially, it is observed that the length of the optimal headline closely mirrors that of other selected headlines. Additionally, we compute the ROUGE scores to compare the optimal headlines against others selected. We use the following two instances for illustrate our observations.

Consider the following headlines that garnered the most votes:

- Dow Falls 64 Points, Comes Within Half a Point of 20K
- Dow Stocks Soar but Fail to Reach 20,000 Mark

Headlines receiving one vote include:

- Dow Nears 20K, But Loses Momentum
- Dow Comes Within Half a Point of 20K
- Dow Closes Below 20K
- Dow Falls Short of 20K

This analysis reveals that while all headlines convey accurate information, their level of informativeness varies. For instance, the first headline specifies a 64-point decline, a detail absent in other titles.

Another noteworthy example is the headline "NBA Season Cancellations Likely to Extend Through November 28 Due to Salary," compared with:

- NBA Season in Jeopardy as Owners Push for 50-50 Revenue Split
- NBA Season Could Be Canceled Through Nov. 28
- NBA May Cancel 2 More Weeks of Season
- NBA to Cancel 2 More Weeks of Season
- NBA Canceling 2 More Weeks of Games? 102 More Games Gone
- NBA Planned to Ax 102 More Games

In this scenario, the optimal headline succinctly conveys the cause (salary), consequence (game cancellations), and timeframe (through Nov. 28), whereas others mention only one or two of these elements. These examples, alongside our statistics, illustrate that brevity does not necessarily equate to superiority. A headline that encapsulates the most crucial information is often more valuable. Consequently, a further proposed open research question for future studies concerns the estimation of the informativeness of the generated headline.

6 Conclusion

In this paper, we explored the complexities of numerical understanding and generation in text, an area that has garnered increasing interest within the NLP community. By introducing and evaluating a set of tasks across diverse datasets, our work highlighted significant progress towards enhancing models' numerical comprehension and their application in practical scenarios, including quantitative analysis and numeral-aware headline generation. Our comprehensive evaluation, encompassing both automatic and human assessments, demonstrated the capabilities and limitations of current methodologies, emphasizing the sophisticated understanding necessary to effectively manipulate and interpret numerical information in textual formats. As we approach the mastery of simple numerical questions with the appropriate selection of language models, our research indicates a shift towards more intricate and nuanced challenges in numerical NLP. The advancements facilitated by NumEval set the stage for future investigations into the deeper integration of numeracy and language, aiming not only for models that comprehend numbers but also for those capable of reasoning, inferring, and generating text that accurately reflects the quantitative dimensions of the world.

Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-112L900901. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Sankalp Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. Noot Noot at SemEval-2024 Task 7: Numerical reasoning and headline generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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. 2023. Improving numeracy by input reframing and quantitative pre-finetuning task. In *Findings of the Association for Computational Linguistics: EACL 2023 (Forthcoming)*. Association for Computational Linguistics.

Kaiyuan Chen, Jin Wang, and Xuejie Zhang. 2024. YNU-HPCC at SemEval-2024 Task 7: Instruction fine-tuning models for numerical understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Hao-Yun Chuang and Ali Zhunis. 2024. Challenges at semeval 2024 Task 7: Contrastive learning approach on numeral-aware language generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Hinoki Crum and Steven Bethard. 2024. hinoki at SemEval-2024 Task 7: Numeval task 3: Numeral-aware headline generation (english). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

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.

Yuming Fan, Dongming Yang, and Xu He. 2024. CTYUN-AI at SemEval-2024 Task 7: Boosting numerical understanding with limited data through effective data alignment. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Andres Gonzalez, Md Zobaer Hossain, and Jahedul Alam Junaed. 2024. Numdecoders at SemEval-2024 Task 7: Flant5 and gpt enhanced with cot for numerical reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

JiangLong He, Saiteja Tallam, Srirama Nakshathri, Navaneeth Amarnath, Pratiba KR, and Deepak Kumar. 2024. Infrrd.ai at SemEval-2024 Task 7: Rag-based end-to-end training to generate headlines and numbers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. NumHG: A dataset for number-focused headline generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. SemEval-2023 Task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Mahmood Kalantari, Mehdi Feghhi, and Taha Khany Alamooti. 2024. MAMET at SemEval-2024 Task 7: Supervised enhanced reasoning agent model. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Tsz-Yeung Lau and Shih-Hung Wu. 2024. CYUT at SemEval-2024 Task 7: A numerals augmentation and feature enhancement approach to numeral reading comprehension. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xinyue Liang, Jiawei Li, Yizhe Yang, and Yang Gao. 2024. Bit_numeval at SemEval-2024 Task 7: Enhance numerical sensitivity and reasoning completeness for quantitative understanding. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Xinyi Liu, Xintong Liu, and Hengyang Lu. 2024. JN666 at SemEval-2024 Task 7: Numeval: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. 2023. SemEval-2023 task 6: LegalEval - understanding legal texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2362–2374, Toronto, Canada. Association for Computational Linguistics.

Zhen Qian, Xiaofei Xu, and Xiuzhen Zhang. 2024. ZXQ at SemEval-2024 Task 7 fine-tuning gpt-3.5-turbo for numerical reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Pawan Rajpoot and Nut Chukamphaeng. 2024. Team NP_PROBLEM at SemEval-2024 Task 7: Numerical reasoning in headline generation with preference optimization. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

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.

Samiran Saha. 2024. Ju united at SemEval-2024 Task 7: Predicting numeral using fine-tuned bert models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Partha Sarathi Sengupta, Sandip Sarkar, and Dipankar Das. 2024. HIJLI_JU at SemEval-2024 Task 7: Enhancing quantitative question answering using fine-tuned bert models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Monika Singh, Sujit Kumar, Tanveen, and Sanasam Ranbir Singh. 2024. ClusterCore at SemEval-2024 Task 7: Few shot prompting with large language models for numeral-aware headline generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vishruth Veerendranath, Vishwa Shah, and Kshitish Ghate. 2024. Calc-CMU at SemEval-2024 Task 7: Pre-calc - learning to use the calculator improves numeracy in language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023.

Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Junzhe Zhao, Yingxi Wang, Huizhi Liang, and Nicolay Rusnachenko. 2024. NCL_NLP at SemEval-2024 Task 7: Cot-numhg: A cot-based sft training strategy with large language models for number-focused headline generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.