Testing the Capability of Small LLMs on SemEval-2024 Task 7: Numeral-Aware Language Understanding and Generation

Deric Liang & Aman Kumar

In our proposed project, we aim to contribute to Task 7 (NumEval), Sub-task 3 (Numerical-Aware Headline Generation) of SemEval 2024. The motivation of the NumEval task is that capturing the semantics of numbers has been difficult in the past. In the summary paper of the NumEval task from Chen et al. (2024), the authors highlight an example where the number significantly impacts the meaning of a sentence: “Stealing $10” versus “Stealing $100,000.”

The specific sub-task is to build systems that generate headlines with an informative number, given a news excerpt. The data for this sub-task is Num-HG, containing news excerpts (features) and journalist-written headlines (labels) outlined in Chen et al. (2024). First, we will test the Numerical Reasoning capability by having the models predict masked numbers from the journalist-written headlines. This is measured by Accuracy (does the number generated in the headline match the journalist headline). Second, we will have the models generate the whole headline. The paper evaluated this task using metrics such as ROUGE, BERTScore, and MoverScore.

Our contribution will come from combining proposed tuning methods with novel models as outlined below:

- Models: Llama 3.2 (1B, 3B), Deepseek-R1-Distill-Llama-8B from Deepseek-AI (2025), GPT-3.5 from Brown et al. (2020), and DeBERTa from He et al. (2020)
- Tuning methods:
    - RAG, He et al. (2024): retrieval-augmented generation
        - Obtains most similar training examples for augmentation to see how other similar examples result in headlines within training data
        - Has the effect of taking a general LLM and enhancing its ability to work on headline generation task specifically
    - Calculator, Veerendranath et al. (2024): Takes in math word problems from MAWPS, SVAMP, and AsDivA to learn context of numbers through arithmetic problems (Citations for data in Veerendranath et al. (2024))
    - Supervised Fine-tuning: Numina-Math 1.5 is a dataset containing 900,000 math problems with their solutions in a Chain-of-Thought manner. This dataset can be used to fine-tune models and effectively teach them how to solve math problems. This skill will likely translate into better performance on the Numerical Headline Generation task.
    - Distillation: Larger Reasoning models such as DeepSeek R1 can be used to produce synthetic examples of mathematical reasoning. Smaller models can be fine-tuned on these examples and consequently improve on numerical understanding and reasoning.

References

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.

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.

Chung-Chi Chen, Jian-Tao Huang, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2024. SemEval-2024 Task 7: Numeral-Aware Language Understanding and Generation. In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). Association for Computational Linguistics.

Deepseek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention.