NumPert:

When Numbers Shift, Does Prediction Hold?

Peter Røysland Aarnes, Vinay Setty
University of Stavanger

Media Futures

Abstract

The accuracy and reliability of Large Language Models (LLMs) in processing numerical information are critical, especially in domains like fact-checking and information verification. This research investigates the vulnerability of Language Models (LMs) to adversarial attacks on numerical claims. By systematically perturbing numerical entities in textual data, we assess how these models handle altered quantitative information. The goal is that our findings could highlight significant weaknesses in current models' abilities to detect and respond to numerical inconsistencies, underscoring the need for enhanced robustness against such adversarial manipulations.

[Model Prediction: TRUE]

[Original Claim] ✅

"In 2020, the company's revenue was **$5 million**, making a significant growth from the previous year."

[Model Prediction: TRUE]

[Perturbed Claim] ❌

"In 2020, the company's revenue was **$50 million**, making a significant growth from the previous year."

[Evidence]

"A market analysis report by MNO Research Group, published in June 2021, states: 'PQR Innovations experienced significant growth compared to the previous year's earnings of $3.8 million. This growth is attributed to successful product diversification and strategic partnerships with (...). The total revenue for the year 2020 reached **$5 million**."

The figure shows that altering numerical claims can mislead models. The original claim of **$5 million** is correctly classified as **True**, but after scaling it to **$50 million**, the model still predicts it as **True**. This highlights a significant vulnerability to adversarial attacks on numerical data and underscores the need for improved robustness in models to evaluate quantitative information accurately.

Research questions

1. How susceptible are current state-of-the-art LLMs models to adversarial attacks involving numerical perturbations in text?
2. What types of numerical perturbations most effectively deceive these models?
3. What strategies are the best to enchance a model's robustness against these types of adversarial attacks?

Background

With the proliferation of misinformation, automated fact-checking systems have become essential tools in combating false narratives. LMs play a pivotal role in these systems by analyzing textual claims and verifying their validity. However, while substantial progress has been made in processing qualitative information, numerical data poses unique challenges. Numerical claims are susceptible to subtle alterations that can drastically change the meaning of a statement, potentially deceiving both models and human readers.
Adversarial attacks exploit these vulnerabilities by introducing carefully crafted perturbations that mislead LMs. Understanding how these models respond to numerical perturbations is crucial for improving their robustness and reliability.

Methods

Dataset: Use True/False claim-evidence pairs from the QuanTemp1 dataset.

Baseline Evaluation: Evaluate the models (GPT 4o, 4o mini, Gemini 1.5 Pro, Llama 3.2 model's, fine-tuned T5 and RoBERTa.

Entity Recognition: Use SpaCy's NER to identify numerical entities within the claims, including quantities, dates, percentages, and monetary values.

Perturbation of claims: Developed scripts to systematically alter numerical entities using various strategies.

- **Numeration Perturbation:** Converting numbers between digits and words.
- **Heterogeneous Number Types**: Changing formatting types and symbols associated with numerical values (e.g. dates, time, fraction, currency symbols).
- **Approximation:** Rounding numbers to nearby values.
- **Range:** Scale measurements (e.g., x value is between y and z).
- **Masking:** Replacing numerical values with placeholder "###".
- **Randomization:** Substituting numbers with random values of the same length.
- **Removal:** Omitting numerical entities entirely.

Assess models performances using perturbed claims: Examine whether the models are susceptible to manipulations of numerical data within the claim component of claim-evidence pairs.

[1]https://github.com/factiverse/QuanTemp

PARTNERS

HOST

UNIVERSITY OF BERGEN

FUNDED BY

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).

Forskningsrådet

Norwegian Centre for Research-based Innovation