NumPert:

Numerical Perturbations to Probe Language Models for Veracity Prediction

Peter Røysland Aarnes, Vinay Setty
University of Stavanger

Media Futures

[Model Prediction: TRUE]

[Original Claim]

"In Q4, the company's revenue was **5,000,000** dollars, making a significant growth from the previous year."

[Paired 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 in Q4 2020 reached **5.000.000** dollars."

[Model Prediction: TRUE]

[Perturbed Claim T → F]

"In Q4, the company's revenue was about **7,500,000** dollars making a significant growth from the previous year."

Figure 1: Example illustrating how the original 'TRUE' claim is perturbed into a 'FALSE' claim, yet the model predicts 'TRUE'.

Abstract

Large language models perform well on fact-checking and question answering but remain weak in numerical reasoning. This study systematically evaluates their robustness in verifying numerical claims using controlled perturbations. Even top models show accuracy drops of up to 62%, with none robust across all conditions. Longer contexts reduce accuracy, but adding perturbed examples helps recovery. Results expose major weaknesses in numerical fact-checking and show that robustness is still an unresolved challenge for current LLMs.

Research question

1. Which models in our selection of diverse sizes are most and least robust?
2. Which numerical perturbations most affect performance?
3. How do context length and reasoning chains influence robustness?

Perturbation Framework

- **Numeration Perturbation:** Converting numbers between digits and words.
- **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.
- **Negative:** Flipping positive percentage value into negative.

- Open-weight models included Llama 3.3-70B, Llama 3.2-1B, Mistral-7B, DeepSeek-R1-32B, and Qwen3-32B (with and without "thinking" variants).

- Proprietary systems comprised GPT-4o, GPT-4o-Mini, GPT-5^{T}, GPT-o3T, and Gemini 2.5 Flash (standard and thinking versions). All models were evaluated in zero-shot, two-shot, and perturbation-aware prompt (PAP) settings on numerical claim–evidence pairs.

- No model proved fully robust, but Gemini 2.5F and Qwen3-32B^{T} emerged as the strongest overall performers. DeepSeek-R1 and smaller Llama variants were the least stable, often leading to performance collapse under two-shot prompts.

- The hardest perturbations were *Masking* and *Negative-Number*, which caused *accuracy drops exceeding 60%,* while Range and Random-Replacement occasionally improved results, indicating models handle approximate or interval values better than precise digits.

- Perturbation-aware prompting improved robustness across all systems, especially reasoning-enabled models like GPT-5^{T} and Gemini 2.5F^{T}.

- Across experiments, longer context and reasoning chains correlated with higher error rates, suggesting that overextended reasoning ("overthinking") leads to misclassification even on simple numerical claims.

Conclusion

Our results show that even leading systems suffer sharp performance drops under controlled numerical edits, providing the first comprehensive evidence that numerical robustness in long-context fact-checking remains an open challenge. Beyond prior work on textual or adversarial perturbations, our study is novel in designing semantically valid numerical perturbations and demonstrating that perturbation-aware prompting can partially recover performance.

PARTNERS

amedia Schibsted 2 VIZRT
NRK Bergens Tidende Faktisk.
nb.no Factiverse WOLFTECH
NORCE dn media group

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