NumPert: Numerical Perturbations to Probe Language Models for Veracity Prediction

Peter Røysland Aarnes **Vinay Setty**
University of Stavanger
peter.r.aarnes@uis.no, vsetty@acm.org

Abstract

Large language models show strong performance on knowledge intensive tasks such as fact-checking and question answering, yet they often struggle with numerical reasoning. We present a systematic evaluation of state-of-the-art models for veracity prediction on numerical claims and evidence pairs using controlled perturbations, including label-flipping probes, to test robustness. Our results indicate that even leading proprietary systems experience accuracy drops of up to 62% under certain perturbations. No model proves to be robust across all conditions. We further find that increasing context length generally reduces accuracy, but when extended context is enriched with perturbed demonstrations, most models substantially recover. These findings highlight critical limitations in numerical fact-checking and suggest that robustness remains an open challenge for current language models.

1 Introduction

Verifying claims in social media, political debates, and press releases has become essential. While platforms such as Politifact, Snopes, and FullFact support manual fact-checking, their scalability is limited. Numerical claims, in particular, are tedious and error prone for human annotators (Aly et al., 2021). Neural language models provide a promising alternative for evidence retrieval and preliminary veracity assessment (Guo et al., 2022; Dmonte et al., 2024; Setty, 2024). Yet, recent studies show that both transformer models fine-tuned for numerical claim verification and general purpose large language models struggle with numerical reasoning (Wallat et al., 2024; V et al., 2024; Akhtar et al., 2023), and the reasons remain unclear.

Numerical Perturbation Example

Original Claim: *"In 2020, the company's revenue was 5,000,000 dollars, making a significant growth from the previous year".*
[**Label:** `TRUE`, **Model Prediction:** **`TRUE`** ✓]

Perturbed Claim: *"In 2020, the company's revenue was fifty million dollars making a significant growth from the previous year."*
[**Label**: `FALSE`, **Model Prediction:** **`TRUE`** ✗]

Evidence: *"A market analysis by MNO Research Group, published in 2021, states: 'PQR Innovations experienced significant growth (...). The revenue for the year 2020 reached 5,000,000 dollars."*[1em]

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

Although prior work has studied LLM fragility in numerical reasoning for QA (Xu et al., 2022) and tabular NLI (Akhtar et al., 2023), no systematic analysis exists for veracity prediction in long-context fact-checking. Our results indicate that models are prone to errors with longer context and reasoning chains. To address this gap, we evaluate state-of-the-art models of different sizes and architectures

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 78–95
December 20-24, 2025 ©2025 Association for Computational Linguistics

under varied prompting settings with systematically perturbed numerical claims and evidence.

Manipulating numerical values in unstructured text requires care to ensure that perturbations remain meaningful. We define six probe types: Numeration (*Num*), Approximation (*Approx*), *Range*, Masking (*Mask*), Random Replacement (*Rand-Repl*), and Negative Number (*Neg-Num*) (see Table 1) to systematically modify numbers while preserving claim intent. In some cases, these perturbations also flip the factual label (e.g., changing $5,000,000 to fifty million; see Figure 1). All perturbations are manually verified to ensure correctness and relevance. This study addresses three research questions:

RQ1 : *Which models in our selection of diverse sizes are most and least robust?*

RQ2 : *Which numerical perturbations most affect performance?*

RQ3 : *How do context length and reasoning chains influence robustness?*

To answer this, we test models on claim–evidence pairs, comparing baseline predictions with those on numerically perturbed claims. Larger gaps reflect weaker robustness. We use truthful probes that keep the original label and label-flipping probes that contradict the evidence, under zero-shot, two-shot, and perturbation aware prompts (PAP).

Our results show that all state-of-the-art models are highly vulnerable to numerical perturbations, particularly under *Mask* and *Neg-Num*. We also notice that zero-shot settings outperform two-shot, while providing a few perturbed examples (PAP prompt) helps models recover in most cases. These findings reveal weaknesses in LLM veracity prediction.

2 Related work

The interpretability of LLMs is critical for knowledge-intensive tasks like question answering and fact-checking. Probing studies have revealed their opaque decision processes (Belinkov, 2022). For instance, Yang et al. (2024); Lu et al. (2023); Frieder et al. (2024) showed that while LLMs can perform complex reasoning, they often struggle with basic numeracy.

Several works have examined numerical reasoning in LLMs. Wallace et al. (2019) probed embeddings from BERT and GloVe, finding inherent but inconsistent numeracy. Akhtar et al. (2023) evaluated models on tabular data with a hierarchical taxonomy, showing no model excels across all tasks. Xu et al. (2022); Zhou et al. (2024) demonstrated that numerical perturbations in QA often mislead LLMs, while Paruchuri et al. (2024); Chen et al. (2024) highlighted weaknesses in numerical reasoning. Several studies also reveal that LLMs for fact-checking are brittle to textual perturbations, and adversarial edits (Mamta and Cocarascu, 2025; Przybyła et al., 2024; Liu et al., 2025).

Despite prior advances, key gaps remain. Most work does not examine numerical reasoning in *open-domain fact-checking* with real-world, long-context data, and reproducibility is often limited. For instance, Akhtar et al. (2023) rely on synthetic tabular inputs with short context, provide incomplete perturbation details, and lack an accessible repository. In contrast, we evaluate numerical reasoning in realistic, unstructured settings, introduce perturbations that preserve semantic validity, and release full code and data to ensure reproducibility.

3 Methodology

Our study examines veracity prediction models by systematically perturbing numerical values in claims to assess their impact on label prediction. The methodology involves (1) curating a dataset with diverse numerical expressions (e.g., statistics), (2) applying controlled perturbations (e.g., scaling, replacements, masking). (3) Extensive error analysis leveraging the reasoning tokens.

Table 1: The number of claims per perturbation type that remain 'True' (T→T), remain 'False' (F→F), or switch from 'True' to 'False' (T→F). Unperturbed baseline has 260 `True` claims and 604 `False` claims.

Category	T → T	F → F	T → F
Num	213	490	213
Approx	170	404	170
Range	188	411	188
Mask	✗	490	213
Rand-Repl	✗	490	213
Neg-Num	✗	89	51

3.1 Dataset and Preprocessing

We use the *QuanTemp* dataset (V et al., 2024), which contains real world claim-evidence pairs with numerical focus from reputable fact checking sources. Each pair is labeled as `True`, `False`, or `Conflicting`. For our evaluation, we exclude the *Conflicting* class due to its inherent ambiguity. To prevent shortcut learning, we remove summaries from all pairs, requiring models to assess veracity solely from evidence.

Each claim is processed with the spaCy NER tagger (covering *Cardinal*, *Money*, *Percent*, *Time*, *Date*, and *Ordinal*), and numerical values are normalized to digits using the *Word2Number* library (similar to (Akhtar et al., 2023; Wallace et al., 2019; Xu et al., 2022)). Perturbed claims are *manually verified* for validity, and invalid cases are removed.

3.2 Perturbation Techniques

We adopt the numerical reasoning taxonomy of Akhtar et al. (2023) (see Table 1). The *Num*, *Approx*, and *Range* settings perturb numbers while remaining consistent with the evidence, so `True` claims stay `True`. Conversely, *Mask*, *Rand-Repl*, and *Neg-Num* modify values such that `True` claims flip to `False`, while `False` claims remain unchanged. We do not perturb `False` to `True`, since falsity can stem from multiple factors and counterfactual claims are often infeasible. Exploring this direction is left for future work. Now we explain the different perturbation techniques:

***Num*:** Tests whether models recognize equivalence between digits and words (e.g., "12" vs. "twelve"), preserving the original label for non-flipping probes. Perturbation applies to `Cardinal`, `Percent`, and `Money`, but not to `Ordinal`, `Time`, or `Date`, except for cardinal numbers within `Time` (e.g., "24 hours" to "twenty four hours"). For the label-flipping probes, the original number is modified (e.g., "12" could be perturbed to "fifteen").

***Approx*:** Non-flipping probes reduces precision by rounding and adding *about* (e.g., "1,025 dollars" to "about 1000 dollars"), retaining truth when close to the evidence. For the label-flipping probes, the original value is altered so that it is no longer reflective of the true amount (e.g., original"1,025 dollars" to "about 1200 dollars").

***Range*:** Non-flipping probes replaces exact values with spans (e.g., "25 percent" to "between 20 and 30 percent"), testing reasoning over intervals. The label-flipping probes modifies the span such that the original number is not within it (e.g., the original "25 percent" is perturbed to "between 30 and 40 percent").

***Rand-Repl*:** Replaces numbers with random values of equal digit length (e.g., "100,000" to "423,823"), mismatching the evidence.

***Mask*:** Hides numbers with "#" tokens according to digit length, including delimiters (e.g., "100,000" to "#######"), requiring inference from evidence.

***Neg-Num*:** Converts values to negatives (e.g., "4%" to "-4%"), applied only to percentages since other entities (money, time, dates) typically use linguistic cues like "decrease."

3.3 Prompting Strategy

All models use identical instructions under three prompting strategies: (1) **Zero-shot** with only instructions and no demonstrations (see Appendix B), (2) **Two-shot** prompt that extends the zero-shot prompt with one `True` and one `False` demonstration from training data with evidence and rationale (Brown et al., 2020). (3) We also test models with a perturbation aware prompt **(PAP)**, which pairs a perturbed claim with one sentence evidence for each perturbation type and flipped label. A similar approach is used by (Hu et al., 2024) in a RAG setting. Full prompts are provided in Appendix B.

4 Experimental Setup

This section describes our experimental framework, including the language models used, and evaluation methods.

4.1 Model Selection

Open-weight LLMs: *DeepSeek-R1-32B*, *Qwen3-32B*, *Llama3.3-70B*, *Llama 3.2-1B*, and *Mistral-7B* (All models are from Ollama framework[1] with Q4_K_M quantization).

Proprietary LLMs: *GPT-4o (v2024-08-06)*, *GPT-4o-mini (v2024-07-18)*, *GPT-5 (v2025-08-07)*, *GPT-o3 (v2025-04-16)*, and *Gemini 2.5 Flash (v2025-06)* (All models are accessed via their respective official APIs)

Models with thinking are marked with superscript T. All models ran with temperature 0 and JSON output; open-weight and OpenAI models used default (medium) reasoning effort. For Gemini 2.5 Flash^T, we fixed the thinking budget to 8192 (vs. the default 1) for cost efficiency. Other settings followed defaults. We exclude *Llama 3.2-1B* and *Mistral-7B* from the main results due to limited robustness; details are in Appendix A.2. Invalid predictions are rare, except for DeepSeek-R1^T, which yields 6.8% invalid outputs under zero-shot. Thinking variants generally produce more invalid outputs than their non-thinking counterparts (see Appendix C). Code and data can be accessed though our GitHub repository[2].

4.2 Evaluation

Robustness is assessed by comparing baseline performance on non-perturbed claims with performance on perturbed ones. We use per-class accuracy metric. We use accuracy as the primary metric for $T \rightarrow F$ evaluations. To gain greater insight into model errors, we manually analyze reasoning tokens of zero-shot vs. PAP for `T` → `F` claims to look for common patterns that models fall into while evaluating a claim.

5 Results

We report results across models and perturbation settings. We first describe performance on unperturbed claims, then analyze changes under non-flipped and flipped label conditions. Results for `False` → `False` cases are omitted here for brevity (see Appendix A.2). Models are evaluated under three prompting regimes defined in Section 3.3 (see Appendix B for full prompts).

5.1 `True` → `False`

We start with the most challenging case: label-flipping perturbations (`True` → `False`), shown in Table 2. Since the claim and ground-truth label are flipped, all reported results reflect the flipped label. A drop in performance means models still predict `True` instead of the expected `False` and less robust. Performance on unperturbed `True` claims is given in the “Original” column as the baseline for each prompting regime.

[1] https://ollama.com/search

[2] https://github.com/iai-group/adversarial_attack_numerical_claims/

Table 2: Accuracy (reported in %) for 'True' dataset split for label flips perturbations (True → False), and comparing accuracy variance between the flipped probes to model performance on unaltered *original* claims accuracy (-x indicates a drop; +x indicates an increase). Values in bold denote the highest accuracy within each perturbation setting, separated by open-weight and proprietary models.

Model	Original	Approx	Neg-num	Num	Rand-repl	Range	Mask
			Zero-shot				
Llama3.3-70B	87.32	87.65$^{+0.32}$	**62.75**$^{-24.58}$	68.54$^{-18.78}$	**91.08**$^{+3.76}$	82.45$^{-4.88}$	10.80$^{-76.53}$
DeepSeek-R1-32B	81.69	**89.41**$^{+7.72}$	39.22$^{-42.47}$	56.34$^{-25.35}$	88.73$^{+7.04}$	81.91$^{+0.22}$	**23.47**$^{-58.22}$
DeepSeek-R1-32B^T	**87.44**	85.06$^{-2.37}$	31.91$^{-55.52}$	69.43$^{-18.01}$	84.73$^{-2.71}$	86.98$^{-0.46}$	10.63$^{-76.81}$
Qwen3-32B	84.35	78.24$^{-6.12}$	43.14$^{-41.21}$	58.78$^{-25.57}$	84.51$^{+0.16}$	80.32$^{-4.03}$	16.43$^{-67.92}$
Qwen3-32B^T	85.99	89.38$^{+3.38}$	34.04$^{-51.95}$	**78.24**$^{-7.75}$	87.88$^{+1.89}$	**87.64**$^{+1.65}$	12.38$^{-73.61}$
GPT-4o	80.00	88.82$^{+8.82}$	47.06$^{-32.94}$	73.24$^{-6.76}$	90.61$^{+10.61}$	91.49$^{+11.49}$	19.25$^{-60.75}$
GPT-4o-Mini	**85.38**	68.24$^{-17.15}$	25.49$^{-59.89}$	56.81$^{-28.58}$	78.87$^{-6.51}$	75.00$^{-10.38}$	11.27$^{-74.12}$
GPT-5^T	76.15	93.53$^{+17.38}$	33.33$^{-42.82}$	**86.38**$^{+10.23}$	89.20$^{+13.05}$	92.02$^{+15.87}$	19.72$^{-56.44}$
GPT-o3^T	75.77	89.41$^{+13.64}$	25.49$^{-50.28}$	84.98$^{+9.21}$	88.73$^{+12.96}$	90.96$^{+15.19}$	21.60$^{-54.17}$
Gemini 2.5F	82.69	**95.29**-$^{+12.60}$	54.90$^{-27.79}$	83.57$^{+0.88}$	**96.71**$^{+14.02}$	**93.09**$^{+10.39}$	**25.82**$^{-56.87}$
Gemini 2.5F^T	71.54	88.82$^{+17.29}$	**58.82**$^{-12.71}$	82.63$^{+11.09}$	89.67$^{+18.13}$	90.43$^{+18.89}$	16.90$^{-54.64}$
			Two-shot				
Llama3.3-70B	**91.55**	72.35$^{-19.20}$	**33.33**$^{-58.22}$	46.48$^{-45.07}$	78.26$^{-13.29}$	57.98$^{-33.57}$	8.92$^{-82.63}$
DeepSeek-R1-32B	89.67	65.29$^{-24.38}$	21.57$^{-68.10}$	37.09$^{-52.58}$	74.70$^{-14.97}$	58.51$^{-31.16}$	12.21$^{-77.46}$
DeepSeek-R1-32B^T	86.32	88.55$^{+2.23}$	22.00$^{-64.32}$	71.15$^{-15.17}$	88.49$^{+2.17}$	87.17$^{+0.85}$	9.43$^{-76.89}$
Qwen3-32B	79.81	70.59$^{-9.22}$	37.25$^{-42.56}$	49.77$^{-30.05}$	66.40$^{-13.41}$	72.87$^{-6.94}$	**20.66**$^{-59.15}$
Qwen3-32B^T	83.49	**86.98**$^{+3.49}$	27.45$^{-56.04}$	**78.20**$^{-5.29}$	**88.76**$^{+5.26}$	**87.23**$^{+3.74}$	12.74$^{-70.75}$
GPT-4o	86.54	82.35$^{-4.19}$	33.33$^{-53.21}$	68.54$^{-17.99}$	87.32$^{+0.79}$	85.64$^{-0.90}$	13.62$^{-72.92}$
GPT-4o-Mini	**89.62**	67.06$^{-22.56}$	27.45$^{-62.16}$	50.70$^{-38.91}$	77.46$^{-12.15}$	73.94$^{-15.68}$	20.19$^{-69.43}$
GPT-5^T	77.69	**91.18**$^{+13.48}$	29.41$^{-48.28}$	84.04$^{+6.35}$	88.26$^{+10.57}$	88.83$^{+11.14}$	18.78$^{-58.91}$
GPT-o3^T	75.77	89.41$^{+13.64}$	23.53$^{-52.24}$	**85.45**$^{+9.68}$	89.67$^{+13.90}$	**90.43**$^{+14.66}$	22.07$^{-53.70}$
Gemini 2.5F	85.00	87.06$^{+2.06}$	35.29$^{-49.71}$	70.89$^{-14.11}$	**94.37**$^{+9.37}$	85.64$^{+0.64}$	**22.90**$^{-62.10}$
Gemini 2.5F^T	74.23	90.00$^{+15.77}$	**52.94**$^{-21.29}$	82.16$^{+7.93}$	92.02$^{+17.79}$	88.83$^{+14.60}$	15.96$^{-58.27}$
			Perturbation Aware Prompt (PAP)				
Qwen3-32B	79.34	89.41$^{+10.07}$	76.47$^{-2.87}$	73.71$^{-5.63}$	90.61$^{+11.27}$	89.36$^{+10.02}$	**67.61**$^{-11.74}$
Qwen3-32B^T	71.23	**95.27**$^{+24.04}$	74.00$^{+2.77}$	**90.14**$^{+18.91}$	**94.37**$^{+23.14}$	**94.62**$^{+23.40}$	44.85$^{-26.38}$
Gemini 2.5F	**81.92**	**97.06**$^{+15.14}$	74.51$^{-7.41}$	84.98$^{+3.05}$	**97.18**$^{+15.26}$	**94.68**$^{+12.76}$	29.11$^{-52.82}$
Gemini 2.5F^T	63.08	91.76$^{+28.69}$	**88.24**$^{+25.16}$	**86.85**$^{+23.78}$	92.02$^{+28.94}$	90.96$^{+27.88}$	26.29$^{-36.79}$

5.1.1 Performance on Unperturbed Claims

In zero-shot, most models cluster in the low to high eighties, with Llama 3.3-70B performing best at about 87% and Qwen3-32B^{T}is close behind at 86%. Proprietary models are slightly lower, with GPT-4o-Mini reaching about 85% as the strongest performer. This suggests that larger models may require more specified prompts to achieve higher accuracy.

With two-shot prompting, baselines increase for Llama 3.3-70B, the GPT variants, and DeepSeek-R1. Llama 3.3-70B surpasses 91%. In contrast, Qwen3-32B variants decline, Gemini 2.5F drops slightly, and its thinking variant shows a modest improvement. Under PAP, both Qwen and Gemini models exhibit performance declines. Models get confused by PAP since it contains counterfactual examples.

Overall, adding few-shot examples improves baselines for Llama and GPT models but tends to reduce them for Qwen and Gemini. No-

Table 3: Accuracy (reported in %) on the 'True' dataset split under non label-flipping perturbations (True → True). The table compares perturbed accuracy to unaltered *original* claim accuracy (-x indicates a drop; +x indicates an increase). Values in bold denote the highest accuracy within each perturbation setting, separated by open-weight and proprietary models.

Model	Approx	Num	Range
Zero-shot			
Llama3.3-70B	71.76$^{-15.56}$	**86.38**$^{-0.94}$	70.21$^{-17.11}$
DeepSeek-R1	75.29$^{-6.40}$	82.63$^{+0.94}$	67.55$^{-14.14}$
DeepSeek-R1^{T}	**81.44**$^{-6.00}$	84.62$^{-2.82}$	**79.23**$^{-8.20}$
Qwen3	73.53$^{-10.82}$	85.88$^{+1.53}$	62.23$^{-22.12}$
Qwen3-32B^{T}	79.39$^{-6.60}$	85.02$^{-0.97}$	78.24$^{-7.76}$
GPT-4o	68.82$^{-11.18}$	80.28$^{+0.28}$	55.32$^{-24.68}$
GPT-4o-Mini	**81.18**$^{-4.21}$	**92.96**$^{+7.57}$	**79.79**$^{-5.60}$
GPT-5^{T}	75.29$^{-0.86}$	77.00$^{+0.84}$	73.40$^{-2.75}$
GPT-o3^{T}	74.71$^{-1.06}$	77.46$^{+1.70}$	77.66$^{+1.89}$
Gemini 2.5F	60.69$^{-22.00}$	79.81$^{-2.88}$	43.92$^{-38.78}$
Gemini 2.5F^{T}	68.24$^{-3.30}$	71.76$^{+0.22}$	61.70$^{-9.84}$
Two-shot			
Llama3.3-70B	84.71$^{-6.84}$	90.14$^{-1.41}$	85.64$^{-5.91}$
DeepSeek-R1	**88.82**$^{-0.85}$	**90.61**$^{+0.94}$	**86.70**$^{-2.97}$
DeepSeek-R1^{T}	82.25$^{-4.07}$	87.50$^{+1.18}$	77.13$^{-9.19}$
Qwen3-32B	72.94$^{-6.87}$	81.69$^{+1.88}$	67.02$^{-12.79}$
Qwen3-32B^{T}	81.66$^{-1.83}$	84.43$^{+0.94}$	77.72$^{-5.77}$
GPT-4o	77.06$^{-9.48}$	85.92$^{-0.62}$	63.83$^{-22.71}$
GPT-4o-Mini	**81.18**$^{-8.44}$	**89.67**$^{+0.06}$	**76.06**$^{-13.55}$
GPT-5^{T}	78.82$^{+1.13}$	80.28$^{+2.59}$	73.94$^{-3.76}$
GPT-o3^{T}	75.29$^{-0.48}$	79.34$^{+3.57}$	74.47$^{-1.30}$
Gemini 2.5F	75.88$^{-9.12}$	87.79$^{+2.79}$	70.74$^{-14.26}$
Gemini 2.5F^{T}	74.12$^{-0.11}$	76.53$^{+2.30}$	71.28$^{-2.95}$
PAP			
Qwen3-32B	58.82$^{-14.39}$	72.74$^{-0.47}$	44.41$^{-28.79}$
Qwen3-32B^{T}	**62.13**$^{-15.46}$	**77.60**$^{+0.02}$	**66.94**$^{-10.65}$
Gemini 2.5F	**60.00**$^{-21.92}$	**81.69**$^{-0.23}$	53.19$^{-28.73}$
Gemini 2.5F^{T}	57.65$^{-5.43}$	63.38$^{+0.30}$	**54.79**$^{-8.29}$

tably, the thinking variants consistently perform slightly worse than their non-thinking counterparts, possibly due to the "overthinking" phenomenon as defined by (Sui et al., 2025), in which reasoning models produce unnecessarily long and elaborate chains of reasoning that ultimately reduce problem-solving efficiency – a pattern confirmed by our error analysis (see Section 6.1). Among open-weight LLMs, performance is stronger in zero-shot and two-shot prompts, but when label-flipping examples are included, Gemini 2.5F outperforms Qwen3-32B.

Performance on unperturbed false claims is generally higher, reflecting the fact that fact-checking tasks predominantly target false claims. Consistent with earlier observations, open-weight models exhibit slightly stronger results than proprietary counterparts. A comprehensive analysis is presented in Appendix A.2.

5.1.2 Performance on Perturbed Claims

Now we summarize the change in performance under numerical perturbation. The Table 2 shows the change in accuracy values in red or green superscript depending on if the accuracy deceases or increases to the corresponding baseline with unperturbed original claims.

Masking and negative number perturbations are consistently the most challenging across prompting regimes. Masking yields very low accuracy in zero-shot setting (max 26%), as models often treat masked tokens as placeholders and predict `True`. With negative numbers, accuracy typically falls below 20% for masking and 30–50% overall, except Llama 3.3-70B, which maintains 63%; many models dismiss negatives as typos. Range and approximation perturbations raise accuracy for Qwen, DeepSeek, GPTs (not Mini), and Gemini, showing a preference for approximate over exact values. Numeration perturbations hurt open-weight models (Qwen3-32B, Llama 3.3-70B) but help proprietary systems (GPT-5^{T}, GPT-o3^{T}, Gemini 2.5F), reflecting stronger handling of surface forms.

In two-shot settings, similar trend to zero-shot is observed with slight drop in performance overall. With notable exceptions being DeepSeek-R1, Llama 3.3-70B, and Qwen3-32B drop sharply on approximation, while thinking models, GPT-5^{T}, GPT-o3^{T}, and Gemini 2.5F^{T}, gain on approximate perturbations.

For the rest of the perturbations, a similar trend to that of zero-shot is observed.

Finally, we find that introducing a single label-flipping demonstration for each perturbation type (PAP, shown in Appendix B) substantially boosts performance across all perturbations. The most striking gains appear in reasoning-oriented models, which display far greater robustness than their non-thinking counterparts. In the case of *Neg-Num*, these models not only surpass their baselines but also achieve strong improvements on perturbations such as simple numeration and ranged replacements. Notably, Qwen3-32B recovers to over 67%, underscoring the effectiveness of this model to leverage perturbed demonstrations, although masking remains a persistent challenge for Gemini. For Qwen, enabling the thinking variant consistently strengthens performance in most cases, whereas for Gemini the benefits are more uneven—showing improvements in certain perturbations but minimal change in others.

5.2 `True` → `True`

Table 3 shows the results for `True` → `True` perturbations. *Neg-Num*, *Rand-Repl* and *Mask* are not relevant when preserving labels.

With few exceptions, most models struggle on *Approx* and *Range* perturbations, though the drop is modest compared to `True` → `False` setting. This suggests that replacing numerical values with approximations or ranges, while preserving truth, can still mislead models into predicting `False`. In contrast, performance under *Num* perturbations remains relatively robust. Unlike label-flipping cases, perturbed PAP does not improve performance; instead, they often confuse models into misclassifying `True` claims as `False`. Surprisingly, GPT-4o-Mini, despite being smaller performs the best under this setting.

6 Discussion

RQ1: Across all experiments, *no single model emerges as universally the most robust*, though Gemini 2.5F and Qwen3-32B models come closest. Our results show that models are generally more robust on `False` claims (Tables 5 and 4) than on `True` claims (Tables 2 and 3). With perturbed false demonstrations, Gemini 2.5F^T achieves near-ceiling accuracy on *Approx*, *Range*, and *Rand-Repl*, and shows the largest recovery on *Neg-Num*; without such calibration, Gemini 2.5F offers the best default balance, consistently leading on *Rand-Repl* and *Range*.

Among open-weight systems, Qwen3-32B^T is the most stable across regimes and uniquely strong on *Mask* when provided perturbed examples, while Llama 3.3-70B excels on zero-shot *Neg-Num* but becomes brittle under two-shot. By contrast, DeepSeek-R1 is the least stable, showing sharp two-shot degradations on *Approx* and *Num*, indicative of harmful anchoring effects.

RQ2: *Neg-Num* and *Mask* appear to be the hardest perturbations among all prompt settings. With perturbation aware prompt (PAP), there is modest recovery and even then the gains are model-dependent (e.g., Gemini 2.5F^T). The *Rand-Repl* and *Range* perturbations are the most straightforward, consistently improving accuracy across models and prompting regimes. The *Num* and *Approx* perturbations fall in the middle: "thinking" models such as GPT-5^T, GPT-o3^T, and Gemini 2.5F^T often gain from these perturbations, while many open-weight base models lose accuracy under two-shot prompts, likely because demonstrations with different numerical notation confuse the models—suggesting that these rely more heavily on superficial formatting cues, making them more sensitive to inconsistencies in numeric representation.

RQ3: Across both Gemini 2.5F^T and

Qwen3-32B^T, misclassified instances consistently involve longer inputs than correct predictions. For Gemini 2.5F^T, misclassifications show $\sim$15% more total tokens than correct cases, largely driven by a $\sim$38% increase in reasoning tokens (877 vs. 635 on average). For Qwen, the effect is even stronger: misclassified examples carry $\sim$41% more total tokens, with reasoning length nearly doubling ($\sim$876 vs. 397, a $\sim$120% increase). Prompt tokens also inflate in misclassifications, albeit more modestly (e.g., $\sim$3–10% increases across models). Taken together, these findings suggest that models tend to fail when they have longer prompt and reasoning tokens (*overthinking* (Sui et al., 2025)), with inflated reasoning chains being a strong marker of misclassification. While PAP prompts introduce longer inputs overall, they provide targeted demonstrations that help mitigate these failures by guiding models toward more stable reasoning. Detailed breakdowns are presented in Appendix B.5.

6.1 Error Analysis

To better understand model errors, we analyze thinking tokens under the $T \rightarrow F$ setting for Qwen3-32B^T and Gemini 2.5F^T, focusing on zero-shot errors that recover in PAP. Appendix C, Table 10 shows specific samples. Our analysis reveals the following reasoning patterns:

Numerical strictness: In PAP reasoning, models tend to interpret numbers more rigidly than in zero-shot. For instance, a claim citing \$330,000 against evidence of \$300,000 was treated as a minor discrepancy in zero-shot, but as a significant mismatch in PAP, predicting `False`.

Masking fallacies: In the zero-shot setting, masked numbers were often treated as placeholders, leading the model to "complete" the claim from evidence rather than verify it. Under PAP reasoning, the model more frequently flagged missing values as critical, aligning with the masked prompt examples and rejecting unverifiable claims. In some cases, however, it ignored the masking and reached the correct verdict, but for spurious reasons such as assuming small discrepancies in the evidence.

Typo interpretation: In the negative-number perturbation setting, under zero-shot, models often interpreted the negative sign (–) as a typo, treating it as a misplaced hyphen and discarding it during evaluation, which led to misclassifications. Under PAP prompting, however, the model highlighted the negative sign as a crucial discrepancy, correctly identifying it as evidence that invalidated the claim.

Overthinking: In some cases, models generate unnecessarily elaborate reasoning that obscures straightforward evidence. For example, for the claim *"Of the [more than 2 million] work opportunities created, more than 1 million have been taken up by the youth"*, the evidence clearly shows 2.5 million created and 1.1 million taken by youth (45%). Instead of rejecting the claim directly, the model speculated about time windows and approximation thresholds, leading to a wrong verdict. This illustrates how excessive reasoning can derail simple numerical checks.

7 Conclusion and Future Work

We introduced a framework for systematically perturbing numerical claims in claim–evidence pairs to evaluate the robustness of state-of-the-art LLMs in veracity prediction. 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.

As a preliminary step, this work opens several directions: perturbing the evidence side of claim–evidence pairs, designing fine-grained probes that target sub-claims, and extending the framework to multi-hop reasoning and counterfactual scenarios.

8 Limitations

Our experiments are constrained by the selection of models tested. Additionally, they were conducted in a black-box environment, restricting access to model weights, parameters, and other internal insights. Some perturbation datasets are also limited in size; a larger and more diverse sample would enhance the robustness of our findings. For reasons discussed in previous sections, our experiments focus exclusively on binary veracity classification ('True' and 'False'), omitting more granular classifications and False-to-True perturbations. Expanding the scope to include these aspects could offer a more comprehensive understanding of model performance under different conditions. Lastly, as with most classification tasks involving LLMs, there is a potential risk of data leakage from training data, which could influence the final evaluation and affect the results.

9 Ethical Considerations

Our research highlights the strengths and weaknesses of various models in binary veracity and counterfactual classification. While this type of research presents valuable opportunities to enhance model security and resilience. However, it also necessitates a thoughtful approach to ethical concerns. For our experiments, some models outperform others, yet we do not endorse any specific model for fact-checking tasks. Fact-checking itself is a nuanced and complex issue. Journalists, fact-checkers, and researchers alike risk introducing inadvertent bias into their work, a concern that also extends to the use of LLMs.

Additionally, while the goal of our experiments is to bring greater attention to LLM performance in specific tasks, these findings also highlight vulnerabilities and encourage the development of more robust models. However, these techniques have multipurpose potential and could be exploited for harmful purposes if misapplied.

Acknowledgements

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

References

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, ACL '23, pages 15391–15405.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Kejie Chen, Lin Wang, Qinghai Zhang, and Renjun Xu. 2024. Metarulegpt: Recursive numerical reasoning of language models trained with simple rules. *arXiv preprint arXiv:2412.13536*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1119–1130.

Fanzhen Liu, Alsharif Abuadbba, Kristen Moore, Surya Nepal, Cecile Paris, Jia Wu, Jian Yang, and Quan Z Sheng. 2025. Adversarial attacks against automated fact-checking: A survey. *arXiv preprint arXiv:2509.08463*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631.

Mamta and Oana Cocarascu. 2025. Facteval: Evaluating the robustness of fact verification systems in the era of large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Akshay Paruchuri, Jake Garrison, Shun Liao, John B. Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11712–11733, Miami, Florida, USA. Association for Computational Linguistics.

Piotr Przybyła, Alexander Shvets, and Horacio Saggion. 2024. Verifying the robustness of automatic credibility assessment. *Natural Language Processing Journal*.

Vinay Setty. 2024. Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, pages 2842–2846.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Preprint*, arXiv:2503.16419.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sigir '24, pages 650–660.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 683–692.

Jialiang Xu, Mengyu Zhou, Xinyi He, Shi Han, and Dongmei Zhang. 2022. Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, EMNLP '22, pages 7950–7966.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2024. Number cookbook: Number understanding of language models and how to improve it. *arXiv preprint arXiv:2411.03766*.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2024. Freb-tqa: A fine-grained robustness evaluation benchmark for table question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2479–2497.

A Appendix

The appendix includes additional details of the perturbation methods used, a summary of the False $\rightarrow$ False evaluation, and the evidence document and evaluation for the two-shot examples.

A.1 Perturbation Details

This section provides a brief description of additional details regarding the perturbation methods. For full script details, refer to the GitHub repository[3].

A.1.1 Numeration

For numbers that should not match the original numerical value in the unperturbed claim, the value is increased by 10%, then converted from digits to words.

A.1.2 Approximation

Each type applies context-specific rounding to create conversational approximations rounding, and adds “about” as an approximation prefix. If all numbers, if it is less than 10 and a decimal number, the number gets round to the nearest .5.

- *Cardinal*: Rounds to tens, hundreds, thousands, or hundred-thousands based on magnitude.
- *Percentage*: Rounds to tens or hundreds, preserving exact values for small percentages.
- *Money*: Similar to Cardinal—with a currency symbol and preserves decimal detail for small amounts.
- *Date*: Rounds to the nearest decade.
- *Time*: Rounds to tens or hundreds depending on magnitude.

For the label-flipping probes, the original numerical value is multiplied randomly by a factor 0.5, 0.6, 1.4, or 1.5, and then rounded as described above.

[3] https://github.com/iai-group/adversarial_attack_numerical_claims/

Table 4: Accuracy performance for the ‘False’ class, in the‘False’ dataset split with perturbations where numerical values have been adjusted to remain similar to the original false claim while maintaining the label, i.e., False → False (-x indicates a drop; +x indicates an increase). Values in bold denote the highest accuracy within each perturbation setting, separated by open-weight and proprietary models.

Model	Original	Approx	Num	Range
		One-shot		
Llama 3.2-1B	5.71	5.45$^{-0.27}$	6.53$^{+0.82}$	6.81$^{+1.10}$
Llama 3.3-70B	93.67	94.06$^{+0.39}$	93.47$^{-0.20}$	92.21$^{-1.46}$
Mistral-7B	96.53	95.79$^{-0.74}$	96.33$^{-0.20}$	95.62$^{-0.91}$
DeepSeek-R1	**97.14**	**97.28**$^{+0.13}$	**96.94**$^{-0.20}$	96.84$^{-0.31}$
DeepSeek-R1^{T}	95.86	95.56$^{-0.31}$	96.13$^{+0.27}$	94.56$^{-1.30}$
Qwen3-32B	96.12	96.29$^{+0.16}$	96.12	**97.32**$^{+1.20}$
Qwen3-32B^{T}	95.92	94.99$^{-0.93}$	95.90$^{-0.01}$	95.15$^{-0.76}$
GPT-4o	**96.52**	**97.28**$^{+0.75}$	**97.14**$^{+0.62}$	**97.08**$^{+0.56}$
GPT-4o-Mini	93.05	93.32$^{+0.27}$	92.45$^{-0.60}$	93.19$^{+0.14}$
GPT-5	95.20	95.05$^{-0.15}$	96.12$^{+0.92}$	95.13$^{-0.06}$
GPT-o3	95.36	94.06$^{-1.30}$	95.92$^{+0.55}$	94.40$^{-0.96}$
Gemini 2.5F	93.21	95.05$^{+1.84}$	93.88$^{+0.67}$	96.11$^{+2.90}$
Gemini 2.5F^{T}	92.05	90.84$^{-1.21}$	90.69$^{-1.36}$	90.02$^{-2.03}$
		Two-shot		
Llama 3.2-1B	10.00	6.93$^{-3.07}$	10.20$^{+0.20}$	9.73$^{-0.27}$
Llama 3.3-70B	95.92	96.04$^{+0.12}$	95.51$^{-0.41}$	93.19$^{-2.73}$
Mistral-7B	87.76	88.12$^{+0.36}$	88.78$^{+1.02}$	87.10$^{-0.65}$
DeepSeek-R1	95.92	95.79$^{-0.13}$	95.71$^{-0.20}$	96.84$^{+0.92}$
DeepSeek-R1^{T}	96.07	95.73$^{-0.34}$	96.27$^{+0.20}$	94.35$^{-1.72}$
Qwen3-32B	**97.76**	**97.28**$^{-0.48}$	**97.76**	**97.32**$^{-0.43}$
Qwen3-32B^{T}	95.91	96.04$^{+0.13}$	96.33$^{+0.42}$	94.88$^{-1.03}$
GPT-4o	**96.36**	**97.28**$^{+0.92}$	**96.12**$^{-0.24}$	**97.32**$^{+0.97}$
GPT-4o-Mini	92.38	95.79$^{+3.41}$	93.88$^{+1.49}$	95.38$^{+2.99}$
GPT-5	95.20	94.55$^{-0.64}$	96.12$^{+0.92}$	95.13$^{-0.06}$
GPT-o3	94.87	94.55$^{-0.31}$	95.10$^{+0.23}$	94.89$^{+0.02}$
Gemini 2.5F	92.72	95.54$^{+2.83}$	94.49$^{+1.77}$	95.62$^{+2.91}$
Gemini 2.5F^{T}	92.38	91.58$^{-0.80}$	94.08$^{+1.70}$	90.27$^{-2.12}$
		PAP		
Qwen3-32B	96.12	**96.78**$^{+0.66}$	**96.53**$^{+0.41}$	**97.20**$^{+1.08}$
Qwen3-32B^{T}	**96.72**	96.40$^{-0.32}$	96.39$^{-0.33}$	95.84$^{-0.89}$
Gemini 2.5F	92.88	**95.30**$^{+2.42}$	92.24$^{-0.64}$	**95.13**$^{+2.25}$
Gemini 2.5F^{T}	**93.54**	90.84$^{-2.70}$	**92.65**$^{-0.89}$	90.02$^{-3.52}$

A.1.3 Range

In the range perturb setting, for when the numerical values should be within the span of the original, the lower bounds we perturb the number by ±10%. For *ordinal*, we subtract and add 1 to the original value to create the range bound.

In instances where the labels are flipped, the numerical span will be outside of the range of the original number.

Table 5: Accuracy performance for the `False` class, in the'False' dataset split with perturbations where numerical values have been modified to differ from the original false claim while preserving the label, i.e., `False` → `False` (-x indicates a drop; +x indicates an increase). Values in bold denote the highest accuracy within each perturbation setting, separated by open-weight and proprietary models.

Model	**Original**	**Approx**	**Neg-num**	**Num**	**Rand-repl**	**Range**	**Mask**
Zero-shot							
Llama2-1B	5.71	5.45$^{-0.27}$	5.62$^{-0.10}$	6.33$^{+0.61}$	4.90$^{-0.82}$	5.35$^{-0.36}$	5.92$^{+0.20}$
Llama3.3-70B	93.67	96.53$^{+2.86}$	93.26$^{-0.42}$	96.12$^{+2.45}$	96.73$^{+3.06}$	95.62$^{+1.95}$	92.86$^{-0.82}$
Mistral-7B	96.53	97.28$^{+0.75}$	95.51$^{-1.02}$	96.33$^{-0.20}$	96.73$^{+0.20}$	96.35$^{-0.18}$	95.31$^{-1.22}$
DeepSeek-R1	**97.14**	**98.51**$^{+1.37}$	**96.63**$^{-0.51}$	97.96$^{+0.82}$	**98.16**$^{+1.02}$	**98.30**$^{+1.15}$	**97.55**$^{+0.41}$
DeepSeek-R1^T	95.86	96.57$^{+0.71}$	91.57$^{-4.30}$	97.42$^{+1.56}$	97.61$^{+1.75}$	97.44$^{+1.58}$	93.74$^{-2.13}$
Qwen3-32B	96.12	98.27$^{+2.14}$	95.51$^{-0.62}$	97.96$^{+1.84}$	97.96$^{+1.84}$	98.30$^{+2.17}$	95.31$^{-0.82}$
Qwen3-32B^T	95.92	96.74$^{+0.83}$	94.32$^{-1.60}$	**98.22**$^{+2.30}$	97.74$^{+1.82}$	98.03$^{+2.11}$	94.01$^{-1.91}$
GPT-4o	**96.52**	**97.77**$^{+1.25}$	**96.63**$^{+0.11}$	**98.16**$^{+1.64}$	**98.16**$^{+1.64}$	**97.81**$^{+1.29}$	**96.53**$^{+0.01}$
GPT-4o-Mini	93.05	96.04$^{+2.99}$	92.13$^{-0.91}$	95.71$^{+2.67}$	95.92$^{+2.87}$	96.84$^{+3.79}$	93.27$^{+0.22}$
GPT-5	95.20	96.29$^{+1.09}$	91.01$^{-4.19}$	97.55$^{+2.35}$	97.35$^{+2.15}$	96.84$^{+1.64}$	95.51$^{+0.31}$
GPT-o3	95.36	96.04$^{+0.68}$	91.01$^{-4.35}$	96.94$^{+1.57}$	96.94$^{+1.57}$	96.84$^{+1.47}$	95.51$^{+0.15}$
Gemini 2.5F	93.21	97.28$^{+4.07}$	94.38$^{+1.17}$	96.94$^{+3.73}$	97.96$^{+4.75}$	97.08$^{+3.87}$	93.88$^{+0.67}$
Gemini 2.5F^T	92.05	93.30$^{+1.25}$	87.64$^{-4.41}$	95.31$^{+3.25}$	94.90$^{+2.84}$	96.09$^{+4.04}$	88.98$^{-3.07}$
2-S							
Llama2-1B	10.00	7.92$^{-2.08}$	7.87$^{-2.13}$	7.76$^{-2.24}$	9.18$^{-0.82}$	9.25$^{-0.75}$	7.76$^{-2.24}$
Llama3.3-70B	95.92	97.52$^{+1.61}$	95.51$^{-0.41}$	96.73$^{+0.82}$	97.87$^{+1.95}$	95.38$^{-0.54}$	95.10$^{-0.82}$
Mistral-7B	87.76	87.38$^{-0.38}$	87.64$^{-0.11}$	88.57$^{+0.82}$	88.78$^{+1.02}$	87.35$^{-0.41}$	87.35$^{-0.41}$
DeepSeek-R1	95.92	98.27$^{+2.35}$	93.26$^{-2.66}$	96.73$^{+0.82}$	85.26$^{-10.66}$	98.05$^{+2.14}$	95.51$^{-0.41}$
DeepSeek-R1^T	96.07	96.50$^{+0.43}$	94.25$^{-1.81}$	97.32$^{+1.25}$	96.79$^{+0.73}$	98.03$^{+1.97}$	94.61$^{-1.46}$
Qwen3-32B	**97.76**	**99.01**$^{+1.25}$	**97.75**$^{-0.00}$	**98.98**$^{+1.22}$	**98.93**$^{+1.18}$	**98.54**$^{+0.79}$	**97.55**$^{-0.20}$
Qwen3-32B^T	95.91	97.52$^{+1.61}$	94.38$^{-1.53}$	97.96$^{+2.05}$	98.04$^{+2.13}$	97.57$^{+1.66}$	96.11$^{+0.20}$
GPT-4o	**96.36**	**98.51**$^{+2.16}$	**98.88**$^{+2.52}$	**97.55**$^{+1.19}$	**97.96**$^{+1.60}$	**98.05**$^{+1.70}$	95.51$^{-0.85}$
GPT-4o-Mini	92.38	96.29$^{+3.90}$	94.38$^{+2.00}$	95.92$^{+3.53}$	96.94$^{+4.55}$	97.08$^{+4.70}$	95.31$^{+2.92}$
GPT-5	95.20	96.04$^{+0.84}$	92.13$^{-3.06}$	97.55$^{+2.35}$	97.35$^{+2.15}$	97.08$^{+1.88}$	**96.12**$^{+0.92}$
GPT-o3	94.87	96.53$^{+1.67}$	91.01$^{-3.86}$	97.35$^{+2.48}$	96.94$^{+2.07}$	97.08$^{+2.21}$	95.31$^{+0.44}$
Gemini 2.5F	92.72	97.03$^{+4.31}$	95.51$^{+2.79}$	96.94$^{+4.22}$	96.73$^{+4.02}$	96.36$^{+3.64}$	93.88$^{+1.16}$
Gemini 2.5F^T	92.38	94.31$^{+1.92}$	89.89$^{-2.50}$	95.94$^{+3.56}$	96.13$^{+3.75}$	96.11$^{+3.72}$	90.82$^{-1.57}$
PAP							
Qwen3-32B	95.10	**98.51**$^{+3.41}$	**95.51**$^{+0.40}$	98.16$^{+3.06}$	98.57$^{+3.47}$	98.54$^{+3.44}$	**97.55**$^{+2.45}$
Qwen3-32B^T	**97.13**	97.77$^{+0.65}$	95.51$^{-1.62}$	**98.98**$^{+1.85}$	**98.57**$^{+1.44}$	**98.54**$^{+1.41}$	95.88$^{-1.25}$
Gemini 2.5F	92.88	**97.52**$^{+4.64}$	**96.63**$^{+3.75}$	**97.14**$^{+4.26}$	**98.16**$^{+5.28}$	**97.81**$^{+4.93}$	**94.29**$^{+1.40}$
Gemini 2.5F^T	**93.54**	94.31$^{+0.76}$	92.13$^{-1.41}$	96.94$^{+3.40}$	96.94$^{+3.40}$	95.62$^{+2.08}$	89.39$^{-4.16}$

A.2 Summary of Model Behavior Under Numerical Perturbations for False Dataset Split (False → False)

Table 5 presents False → False perturbations where numerical values are modified while preserving the false label. Our experiments reveal that large models (e.g., GPT-4o, GPT-4o-Mini, Gemini 2.5F) and open-weight DeepSeek-R1^T maintain high robustness across perturbations, with accuracies typically above 90%. Smaller models such as Llama 3.2-1B and Mistral-7B degrade sharply, especially under *Approx* and *Range*. Qwen3-32B^{T}performs consistently well across shots, rivaling proprietary systems. Notable anomalies include Gemini

Example 1	Example 2
Claim: As Republicans try to repeal the Affordable Care Act, they should be reminded every day that **36,000** people will die yearly as a result.	**Claim:** We see a quarter-billion dollars in a pension fund that needs to be funded at **$1.2 billion**.
Evidence: *Gift Article Share* "As Republicans try to repeal the Affordable Care Act, they should be reminded every day that **36,000 people** will die yearly as a result." — **Sen. Bernie Sanders (D-Vt.)**, in a tweet, Jan. 12, 2017.	**Evidence:** Providence Mayor Angel Taveras had to deal with near bankruptcy in the capital city after he took office in 2011. As the city struggled to fix its budget problems, he won union concessions to reduce pension costs. The most recent figures show the plan is only **31.4-percent funded**.
Evaluation: False	**Evaluation: True**

Table 6: True and False examples of claims and their labels based on evidence used in the prompt.

2.5F's drop under *Approx* (-5 to -6 points) despite strong overall performance, and GPT-4o-Mini's unexpected gains in two-shot (+3 points). Reasoning-enabled (^{T}) variants generally improve robustness, though Gemini's thinking variant remains more variable.

The table 4, reports accuracy metric for the False class in the False dataset split with perturbations. Perturbations significantly modify numerical values while preserving the label (False $\rightarrow$ False). Results are presented for multiple LLMs including Llama, Mistral, DeepSeek, GPT, Gemini, and Qwen across three evaluation setups: Zero-shot, two-shot, and Perturbation-Aware Prompt (PAP). The columns indicate different perturbation types: Original (baseline), Approx, Neg-num, Num, Rand-repl, Range, and Mask. Superscripts with negative values denote drops relative to the baseline, and positive values denote improvements.

In the zero-shot setting, DeepSeek-R1, GPT-4o, and Qwen3-32B^T achieve the highest and most stable performance, maintaining accuracies between 96% and 98% across perturbations. Gemini 2.5F is also stable with scores in the range of 93% to 97%. In contrast, smaller models such as Llama 3.2-1B perform poorly with accuracies around 5–6%. Mid-sized models like Llama 3.3-70B and Mistral-7B perform well but remain slightly below the frontier models.

In the two-shot setting, accuracy improves slightly compared to Zero-shot, especially for the smaller models. DeepSeek-R1 remains strong with scores around 96–97%, GPT-4o reaches 95–98%, Qwen3-32B^T achieves 94–98%, and Gemini 2.5F^T remains consistent with 90–96%. Llama 3.2-1B, however, continues to perform poorly with accuracies only between 7% and 10%.

Perturbation-Aware Prompt (PAP) delivers the highest overall accuracies. Qwen3-32B^T and DeepSeek-R1^T achieve 95–99% across all perturbations, while Gemini 2.5F^T also shows strong performance with accuracies between 89% and 97%. PAP consistently improves the already strong models by about 1–2 percentage points compared to zero-shot and two-shot.

In general, model scale is critical. Small models such as Llama 3.2-1B collapse under this evaluation, while large-scale and frontier models like DeepSeek, GPT-4o, Qwen, and Gemini perform near ceiling. Prompting with two-shot increases stability across most models, and PAP proves to be the most robust method, yielding the best and most consistent results overall.

B Prompt

For the LLMs we use the same instruction and two-shot examples. The zero-shot only includes the instruction, whereas the two-shot includes the instruction and the sample data. The following two-shot examples are snippets of the examples used. For the full prompt, refer to our GitHub repository.

B.1 System Prompt

The following prompt was used as the model system prompt:

You are a professional fact checker, your task is to classify whether the given claim is true or false based on the evidence text provided.

B.2 Instruction

The following prompt was used along with two examples from Table 6:

Given the claim and evidence provided, classify the claim as "label": true if it is true, and "label": false if it is false.

B.3 Two-shot Examples

Table 6 presents two examples of fact-checking claims used in the prompt for LLMs along with their corresponding evidence and veracity evaluations. The two examples are used for all LLMs and all perturbation inputs to be consistent. And each of the two example represents the two distinct labels in the dataset.

B.4 Perturbation Aware Prompt

The following prompt was added to the instruction prompt for the negative example experiments:

The numbers in the evidence may not match the claim. For example:

Claim: The Eiffel Tower is three hundred and fifty-one meters tall. Evidence: The Eiffel Tower is 330 meters tall. "label": false

Claim: The year-over-year U.S. inflation rate at the end of 2024 was -2.9%. Evidence: The year-over-year U.S. inflation rate at the end of 2024 was 2.9"label": false

Claim: The birth rate in Japan in 2023 was between 2 to 2.5. Evidence: The birth rate in Japan in 2023 was 1.2. "label": false

Claim: The population of Canada in 2023 was about 45 million. Evidence: The population of Canada in 2023 was 40.5 million by October 2023. "label": false

Claim: Saturn has 789 moons. Evidence: Discoveries bring Saturn's total moon count to 274, nearly triple Jupiter's and more than the total number of known moons around the other planets. "label": false

Claim: The Wembley Stadium in London has a seating capacity of ######. Evidence: The Wembley Stadium in London has a seating capacity of 90,000. "label": false

B.5 Prompt Length Analysis

We perform prompt length analysis for misclassified instances compared to correct classifications for the two most stable models–Gemini 2.5F^T and Qwen3-32B^T.

Gemini 2.5F^T In misclassified instances, Gemini 2.5-Flash tends to have longer reasoning token length overall, with average total token length increasing by 15% compared to correct predictions (2103 vs. 1822 tokens). Prompt tokens show only a modest difference (+3%). The distribution further suggests that

Perturbation	Prompt Tokens		Reasoning Tokens	
	Misclassified	Correct	Misclassified	Correct
Approximation	2158.7	1303.9	1265.1	371.2
Negative Number	1214.5	1073.2	846.5	339.0
Numeration	1648.2	1239.6	796.1	378.4
Random Replacement	1576.4	1323.8	713.2	363.8
Range	1963.4	1315.1	698.4	401.1
Masking	1234.7	1017.8	717.1	427.7

Table 7: Comparison of average prompt and reasoning token lengths for Qwen3-32B^T between **misclassifications** and **correct classifications** in the Zero-shot setting.

errors are associated with longer and more variable reasoning chains (max reasoning length over 6k tokens), whereas correct predictions are achieved with more compact reasoning. In other words, misclassifications correlate strongly with *overthinking*.

Qwen3-32B^T For Qwen3-32B^T, misclassified cases consistently exhibit inflated reasoning lengths compared to correctly classified instances in the Zero-shot setting (Table 7). For example, reasoning tokens nearly triple in *Approx* (1265 vs. 371) and more than double in *Num* (796 vs. 378) and *Range* (698 vs. 401). Prompt lengths are also consistently higher for misclassifications, with the most pronounced gap in *Approx*, where prompts expand by over 65% (2159 vs. 1304). The anomaly occurs with *Mask*, where reasoning remains high even in misclassifications (717 vs. 428), indicating that masked inputs elicit extended elaboration regardless of correctness. Overall, Qwen3-32B^T tends to over-reason when it misclassifies, while correct predictions are characterized by shorter, more efficient reasoning chains and more compact prompts. All token lengths for Qwen3-32B^T zero-shot settings are shown in Table 7.

Model	Total Instances	Invalid	% Invalid
Zero-shot			
DeepSeek-R1:32B	8841	0	0.00
DeepSeek-R1:32B^T	6837	477	6.98
LLaMA-3.2 1B-Instruct	6837	0	0.00
LLaMA-3.3 70B	6837	0	0.00
Mistral-7B	6837	0	0.00
Qwen-3 32B	7041	0	0.00
Qwen-3 32B^T	6553	165	2.52
Two-shot			
DeepSeek-R1:32B	6951	0	0.00
DeepSeek-R1:32B^T	6951	78	1.12
LLaMA-3.2 1B-Instruct	6837	0	0.00
LLaMA-3.3 70B	6951	0	0.00
Mistral-7B	6837	0	0.00
Qwen-3 32B	6951	0	0.00
Qwen-3 32B^T	6951	23	0.33
PAP			
DeepSeek-R1:32B	6837	0	0.00
DeepSeek-R1:32B^T	6837	92	1.35
LLaMA-3.2 1B-Instruct	6837	0	0.00
LLaMA-3.3 70B	6837	0	0.00
Mistral-7B	6837	0	0.00
Qwen-3 32B	6837	0	0.00
Qwen-3 32B^T	6837	55	0.80

Table 8: Invalid outputs across open-weight models, grouped by shot setting. Thinking-enhanced variants are marked with ^{T}. Percentages are calculated as invalid/total $\times$ 100.

C Invalid Output Analysis

As shown in Table 8, across the open-weight models, invalid outputs are virtually absent in the non-thinking variants: Llama 3.3-70B, Llama-3.2 1B instruct, Mistral-7B, Qwen3-32B, and DeepSeek-R1 consistently produce 0.00% invalidity across all shot settings. By contrast, enabling thinking introduces instability. For instance, DeepSeek-R1^T exhibits a sharp rise in invalid generations under zero-shot (6.98%), which decreases under two-shot (1.12%) and PAP (1.35%), indicating some recovery with examples. Similarly, Qwen3-

Model	Total Instances	Invalid	% Invalid
	Zero-shot		
GPT-4o	5298	0	0.00
GPT-4o-mini	5298	0	0.00
GPT-5	5298	0	0.00
GPT-o3	5298	1	0.02
Gemini-2.5^T	5295	174	3.29
Gemini-2.5	5298	1	0.02
	Two-shot		
GPT-4o	5298	0	0.00
GPT-4o-mini	5298	0	0.00
GPT-5	5298	0	0.00
GPT-o3	5298	2	0.04
Gemini-2.5^T	5298	82	1.55
Gemini-2.5	5298	42	0.79
	PAP		
Gemini-2.5^T	5298	231	4.36
Gemini-2.5	5298	0	0.00

Table 9: Invalid outputs across proprietary models and Gemini variants, grouped by shot setting. Percentages are calculated as invalid/total $\times$ 100.

32B^{T}yields 2.52% invalidity in zero-shot, reduced to 0.33% under two-shot, but climbing again to around 0.80–1.33% with perturbation-aware prompts.

These findings reveal a clear pattern: while open-weight models without reasoning remain robust against invalid outputs, their reasoning-enabled counterparts are consistently more error-prone. Zero-shot is the most unstable regime, two-shot offers stabilizing effects, and PAP shows mixed outcomes depending on the model.

As shown in Table 9, for proprietary systems, invalid outputs are exceedingly rare except in the Gemini family. GPT-4o, GPT-4o-Mini, GPT-5^T, and GPT-o3^{T}maintain near-perfect robustness across all shot settings, with invalid rates between 0.00–0.04%. By contrast, Gemini 2.5F demonstrates a higher susceptibility to invalid generations. In particular, Gemini 2.5F^T reaches 3.29% invalidity under zero-shot, improves to 1.55% with two-shot calibration, but rises again to 4.36% under PAP. Its non-thinking counterpart is much more stable, remaining below 1% invalidity in all cases.

Overall, proprietary models without reasoning capabilities (GPT family and Gemini 2.5F) are extremely stable, while reasoning-enabled variants of Gemini incur significantly higher invalid rates, especially in zero-shot and PAP settings.

True → False Perturbed Claims	**Reasoning Output Snippets**	**Analysis Summary**
The various offices of state government have purchased more than $########## in vehicles, not counting leases, for use from the Governor on down during the last two years.	**Gemini 2.5F**^{T} zero-shot: "(...)While the claim uses a placeholder, the evidence substantiates the 'more than $11 million´figure with precision. I will now mark the claim as true." **Gemini 2.5F**^{T} PAP:"(...)I'm now certain the claim is false. The instruction's emphasis on placeholders makes the 'more than' condition irrelevant. The claim is inherently un-verifiable (...)"	In the zero-shot evaluation the model assumes that the model needs to find the missing value. PAP nudges the model in the correct direction, predicting the claim as false for the correct reasons.
Some states, like Montana and Nebraska, are getting more than three hundred and thirty thousand dollars in federal stimulus money per reported COVID-19 case. New York is the hardest-hit state and yet we are getting only about thirteen thousand, two hundred dollars per case.	**Gemini 2.5F**^{T} zero-shot: "(...) I am also considering whether the small discrepancy in the funding figures makes the overall claim inaccurate (...) The claim is essentially true, even with minor discrepancies" **Gemini 2.5F**^{T} PAP: "(...) claim states '$13,200' per case for New York, whereas the evidence consistently states '$12,000'. Since the numbers don't align, I must classify the claim as false."	PAP highlights that the 10% difference in the claim's numerical values, compared to the evidence as "significant numerical discrepancies". Zero-shot evaluation is more forgiving with regarding the discrepancies.
My opponent, Rick Gunn, blocked the expansion of Medicaid — costing half a million people health insurance, including about 34,000 veterans.	**Qwen3-32B**^{T} zero-shot: "(...)The claim states 34,000 veterans, but the evidence only supports up to 23,000 or 14,000. The user's claim says 34,000, which is higher than both estimates (...)" *(starts to overthink).* **Qwen3-32B**^{T} PAP: "(...) The 34,000 figure isn't present in the evidence; the highest is 23,000. Therefore, the claim is false because the specific number provided doesn't match the evidence."	During zero-shot, the model starts to overthink, going in circles–outputting nearly 7000 reasoning tokens, citing the number in the evidence "23,000", 198 times, and the claim number "34,000", 135 times. During PAP, the model does correctly identify the discrapancy effectively, and keeps the reasoning token output of around 200.

Table 10: Examples of claims, reasoning, and analysis for Gemini 2.5F^T and Qwen3-32B^T where reasoning improves for PAP, compared to zero-shot.