Towards Robust Numerical Claim Verification

Peter Røysland Aarnes
University of Stavanger
peter.r.aarnes@uis.no

Vinay Setty
Factiverse AI and University of Stavanger
vsetty@acm.org

Abstract

Large language models (LLMs) are widely used for claim verification, yet remain brittle for numerical reasoning: even small changes in value can sharply degrade accuracy. We show that this brittleness persists in frontier LLMs, but can be mitigated through adversarial fine-tuning on numerically perturbed examples. Using parameter-efficient fine-tuning, small Qwen3 models (0.6B–8B) reach 98.7% accuracy on label-flipping perturbations, outperforming larger zero-shot models and frontier systems (GPT-5.4 Pro (74.0%) and Gemini 2.5 Flash (73.9%)). The gains generalise to unseen perturbation types, indicating robust numerical decision boundaries rather than memorised edits. Robustness also transfers without target-domain data, significantly improving cross-lingual performance in Spanish. We further show that the same fine-tuning recipe confers robustness to evidence-side perturbations, using the VitaminC dataset.

1 Introduction

Claim verification is an important societal task, and LLMs are a popular choice. Claims that contain numerical values such as statistics, counts, percentages, and dates are especially challenging to verify, even for humans (Aly et al., 2021) and for frontier large language models (V et al., 2024; Aarnes and Setty, 2025; Akhtar et al., 2023), as they require careful interpretation of context (time spans, units, denominators) and cross-checking against reliable sources. Figure 1 illustrates the problem: a simple surface rewrite that changes "3.5%" to "six percent" flips the correct label, yet a zero-shot LLM misses it.

Despite strong performance on arithmetic or structured quantitative tasks (Fang et al., 2025; Schwartz et al., 2024), LLM robustness remains uneven in realistic verification settings. Recent works suggest that even frontier models can break down when the numerical values in a claim are slightly perturbed, revealing brittleness and sensitivity that can undermine trust in their outputs (Akhtar et al., 2023; Aarnes and Setty, 2025; Yang et al., 2025). This brittleness is particularly problematic in evidence-grounded verification, where a model must reconcile a claim's surface form (e.g., a written-out number) with the evidence's surface form (e.g., a digit), and anchoring on lexical overlap rather than semantic equivalence is an easy failure mode. This gap highlights that strong performance on benchmark-style numerical reasoning does not automatically translate into robust, real-world verification of numerical claims.

Numerical Perturbation Example

Original Claim: *"The US unemployment rate fell to 3.5% in February 2020, matching the lowest rate in five decades."*
[**Label:** TRUE, **Zero-shot:** TRUE ✓, **Fine-tuned:** TRUE ✓]
Perturbed Claim: *"The US unemployment rate fell to six percent in February 2020, matching the lowest rate in five decades."*
[**Label:** FALSE, **Zero-shot:** TRUE ✗, **Adversarially Fine-tuned:** FALSE ✓]
Evidence: *"(...) According to the U.S. Bureau of Labor Statistics: 'The unemployment rate was 3.5 percent in February 2020, unchanged from January 2020 (...)' "*

Figure 1: Numeration perturbation: rewriting "3.5%" as "six percent" flips the label to FALSE. Zero-shot predicts TRUE (wrong); adversarial fine-tuning predicts FALSE (correct).

In this paper, we investigate whether adversarial fine-tuning (training on numerically perturbed claims) can improve robustness in numerical claim verification. We apply six perturbation categories (numeration, approximation, range, negation, masking, and random replacement) to the QuanTemp dataset (V et al., 2024), fine-tune Qwen3 models from 0.6B to 8B parameters using QLoRA and soft prompt tuning, and evaluate cross-dataset generalisation to English and Spanish corpora unseen during training, addressing four

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3894–3915
November 6-10, 2026 ©2026 Association for Computational Linguistics

research questions:

RQ1 : *Does adversarial fine-tuning improve robustness to label-flipping perturbations over clean-data fine-tuning?*

RQ2 : *Does training on a perturbation subset generalise, and how does training composition affect robustness?*

RQ3 : *How do model scale and adaptation method (QLoRA vs. soft prompts) interact with adversarial fine-tuning?*

RQ4 : *Does adversarial fine-tuning generalise across domains and languages?*

Our experiments show that adversarial fine-tuning is highly effective: a model trained on clean data alone collapses to under 13% accuracy on label-flipping perturbations, while QLoRA fine-tuning on all six perturbation types restores accuracy to 96–100%. Training composition matters more than variety; fine-tuned small models (0.6B) surpass zero-shot 32B models on the hardest perturbations. Crucially, adversarially fine-tuned Qwen3-8B models also substantially outperform frontier closed-source models (GPT-5.4 Pro, Gemini 2.5 Flash) – for instance, achieving 97.7% vs. 74.0% micro-average accuracy on label-flipping perturbations. Cross-dataset generalisation is practically robust: fine-tuned models outperform the zero-shot baseline on Spanish (97.5% vs. 96.4%), whereas on AVeriTeC the accuracy difference is not significant and the gain is in coverage (the baseline returns no parseable verdict on 37.5% of instances). Detailed answers to each RQ are in Section 5.

Our main contributions are: **(i)** a systematic study of LLM brittleness to numerical claim perturbations and its recovery by fine-tuning on adversarial (counterfactually constructed) examples; **(ii)** ablations showing which perturbation types are *essential* (*Masked*, *Numeration*) and which are redundant (*Approximation*, *Random*); **(iii)** the counterintuitive finding that unperturbed training examples, rather than the perturbations themselves, shift the decision boundary and reduce label-preserving accuracy; **(iv)** evidence that adversarially fine-tuned small models (0.6B–8B) substantially outperform zero-shot 32B and frontier closed-source models (GPT-5.4 Pro, Gemini 2.5 Flash) on label-flipping perturbations; and **(v)** the first cross-lingual evaluation showing numerical robustness transfers from English to Spanish without target-language training data.

2 Related work

Benchmarks such as FEVER (Thorne et al., 2018) and methods from retrieval pipelines to program-guided reasoning (Guo et al., 2022; Pan et al., 2023) are popular. Numerical claims are a particularly hard sub-problem (V et al., 2024; Lu et al., 2023); we build on QuanTemp (V et al., 2024), and the perturbation taxonomy of Aarnes and Setty (2025).

LLM robustness to numerical perturbations is fragile. While targeted training and better encodings reduce numerical errors (Wallace et al., 2019; Geva et al., 2020; Schwartz et al., 2024), small value changes still flip predictions across QA (Akhtar et al., 2023; Xu et al., 2022; Zhou et al., 2024a), claim verification (Aarnes and Setty, 2025; Mamta and Cocarascu, 2025), and mathematical reasoning (Yang et al., 2025). The brittleness is compounded in multilingual settings, where surface-form variation across languages (e.g., decimal separators, numeral systems, written-out forms) creates additional attack surface that monolingual robustness studies have not addressed (Alam et al., 2025).

Adversarial training addresses this brittleness: worst-case perturbation augmentation (Madry et al., 2018), gradient-based NLP perturbations (Guo et al., 2021), and math-problem adversaries (Zhou et al., 2024b) all improve robustness, as do counterfactually revised (Kaushik et al., 2020) and adversarially collected (Nie et al., 2020) training sets. We apply this strategy to numerical claim verification across six perturbation types.

Parameter-efficient fine-tuning makes large-model adaptation practical via LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023), and soft prompt tuning (Lester et al., 2021). Fine-tuned models consistently outperform larger zero-shot LLMs on fact-checking (Setty, 2024; Tang et al., 2024; Larsson et al., 2026). To the best of our knowledge, there are no prior studies applying adversarial fine-tuning for numerical perturbations in claim verification setting.

3 Methodology

Our method consists of two main steps: First, we generate adversarial perturbations for numerical claims, where we adopt a similar taxonomy and methodology to Akhtar et al. (2023); Aarnes and

Category	T→T	T→F	F→F	Total
Numeration	216	216	1002	1434
Approximation	139	139	784	1062
Range	190	190	843	1223
Masked	✗	216	501	717
Random	✗	216	501	717
Negative	✗	52	91	143

Table 1: QuanTemp **test** split: sample counts by perturbation type and transition direction. Original claims: 264 `True`, 615 `False`.

Setty (2025). Second, we conduct adversarial fine-tuning on a set of models, using a mix of perturbed and original data.

3.1 Generating Numerical Perturbations

We use the QuanTemp dataset (V et al., 2024) of real-world numerical claim–evidence pairs, with evidence verdict summaries removed and labels binarised to `True`/`False`.

Dataset and Perturbation Types. Numerical values in claims are automatically tagged by a spaCy[1] named entity tagger model. We apply the six perturbation types defined by Aarnes and Setty (2025) to the numerical values (see Appendix A.2 for definitions and examples): **Numeration** (digit↔word form, e.g. "350"→"three hundred fifty"); **Approximation** (insert "about"); **Range** (exact value → numerical range); **Random** (substitute a random value of equal character length); **Negation** (invert the value); **Masking** (replace the value with "#" symbols of equal character length). For numeration, approximation, and range, we generate both a value-preserving and a value-changing variant, which are label-preserving and label-flipping for `True` claims, while both remain `False` for `False` claims; random, negation, and masking produce only `False` variants. Of note, masking differs from the other types by making a claim false through the omission of its numerical values. To prevent cross-split leakage, perturbations are generated within each predefined QuanTemp split, so perturbed variants of a claim remain in the split of their source. Table 1 reports the statistics for the test split; statistics for the training and validation splits are provided in Appendix A.1.

Cross-Dataset Generalisation Splits. We evaluate generalisation on two held-out datasets: **AVeriTeC 2.0** dev (Schlichtkrull et al., 2023; Akhtar et al., 2025) (119 numerical claims, 631 perturbed instances) and a **Spanish** test set from CLEF CheckThat! 2025 (Venktesh et al., 2025) (2,633 claims, 13,827 perturbed instances). Dataset statistics and construction details are in Appendix A.3.

[1] https://spacy.io/api/entityrecognizer

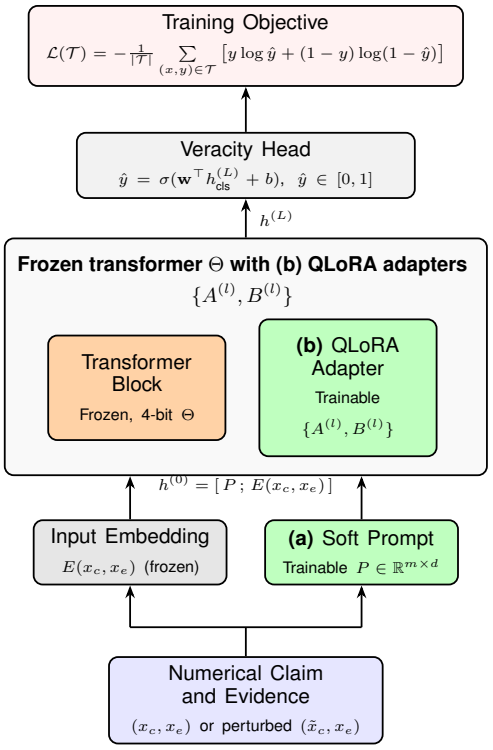


Figure 2: Two PEFT strategies for numerical claim verification. **(a) Soft prompt tuning** prepends m learnable vectors P to frozen embeddings; **(b) QLoRA** inserts low-rank residuals into a 4-bit quantized backbone Θ. Both share a sigmoid veracity head trained on mixture $\mathcal{T}$ of original and perturbed examples.

3.2 Adversarial Fine-tuning

We frame numerical claim verification as binary classification: given a claim x_c and evidence x_e, a model predicts $\hat{y} \in \{\texttt{True}, \texttt{False}\}$. Adversarial fine-tuning augments the standard training set $\mathcal{D}_{\text{orig}} = \{(x_c^{(i)}, x_e^{(i)}, y^{(i)})\}$ with perturbed instances $\mathcal{D}_{\text{adv}} = \{(\tilde{x}_c^{(i,t)}, x_e^{(i)}, \tilde{y}^{(i,t)})\}$, where $\tilde{x}_c^{(i,t)}$ is claim i perturbed under type t and $\tilde{y}^{(i,t)}$ is the revised label. The full training mixture $\mathcal{T} = \mathcal{D}_{\text{orig}} \cup \mathcal{D}_{\text{adv}}$ is optimised under binary cross-entropy:

$$\mathcal{L}(\mathcal{T}) = -\frac{1}{|\mathcal{T}|}\sum_{(x,\,y)\,\in\,\mathcal{T}}\big[y\log\hat{y}+(1-y)\log(1-\hat{y})\big]$$

We compare two independent parameter-efficient adaptation strategies (Figure 2), both keeping the pre-trained backbone Θ frozen.

QLoRA (Hu et al., 2022; Dettmers et al., 2023). The backbone Θ is loaded in 4-bit NormalFloat (NF4) precision. A low-rank residual is inserted into each selected projection matrix $W^{(l)} \in \mathbb{R}^{d\times k}$:

$W^{(l)} \leftarrow W^{(l)} + B^{(l)}A^{(l)}, A^{(l)} \in \mathbb{R}^{r\times k}, B^{(l)} \in \mathbb{R}^{d\times r}$ with rank $r \ll \min(d, k)$. $A^{(l)}$ is initialised from a Gaussian distribution and $B^{(l)}$ is initialised to zero, so $\Delta W^{(l)} = 0$ at the start of training. Only $\{A^{(l)}, B^{(l)}\}$ and the veracity head are trained. Adapters are inserted into all attention (Q, K, V, O) and feed-forward (gate, up, down) projection matrices with rank $r = 32$, implemented via the Unsloth[2] framework.

Soft prompt tuning (Lester et al., 2021) prepends m (set to 20) learnable continuous vectors $P \in \mathbb{R}^{m\times d}$ to the frozen input embeddings, training only P and a veracity head while all backbone weights remain fixed. For the full formulation, see Appendix B.2.

3.2.1 Training Data Sampling

We use four training configurations: (1) original only; (2) leave-one-out (all perturbation types except one); (3) single perturbation type; and (4) all perturbation types. Each is balanced 1:1 between original and perturbed examples; the validation set retains its natural distribution.

3.3 Evaluation and Fine-tuning

A mid-sized base model serves both as the primary model for fine-tuning and as the baseline for different prompting strategies. We also evaluate a larger base model variant and two proprietary models in a zero-shot setting to establish an upper-bound reference point. Lastly, three smaller model variants are fine-tuned with all perturbations and original data for traditional adversarial training (see Section 4 for the specific models used). Full code and data can be accessed through our GitHub repository[3].

3.4 Prompting Strategy

All models use the same system and instruction prompts for fine-tuning and evaluation. Besides the *Zero-shot* setup, the base model is also evaluated with a *Two-shot* strategy, including one `True` and one `False` claim–evidence pair. These prompts are provided in Appendix B.3. We also tested Numerical Counterexample Prompt (NCP), but omit it from the main evaluation due to high parse-failure rates (42–61%; details are found in Appendix C.2).

[2] https://unsloth.ai/

[3] https://github.com/iai-group/numpert-fine-tune

4 Experimental Setup

4.1 Model Set and Inference Settings

We use Qwen3-8B as the primary model, with additional experiments across Qwen3 0.6B, 1.7B, 4B, and 32B (Unsloth Q4_K_M GGUF quantisation (Dettmers et al., 2023)). Models run at 0.001 temperature, with thinking enabled, and use the prompting strategy described in Section 3. The Qwen3 family is selected because it provides openly available checkpoints from 0.6B to 32B under a consistent pre-training recipe, enabling controlled scale ablations within a single family. Aarnes and Setty (2025) report the same numerical brittleness on LLaMA-family models for the underlying claim verification task, so we expect our qualitative findings to reflect the perturbation taxonomy and verification task rather than Qwen3-specific properties; quantitative effect sizes may differ across families. Results for unquantised models can be found in Appendix H.

As closed-source frontier baselines we evaluate $\mathrm{M}_{\text{GPT-5.4}}$ (GPT-5.4 Pro via the Azure OpenAI API, version `2024-12-01-preview`) and $\mathrm{M}_{\text{Gemini-2.5}}$ (`gemini-2.5-flash`, knowledge cutoff January 2025), both queried zero-shot with the same prompt template.

4.2 Parameter-Efficient Adaptation

Models are trained for 2 epochs with a 1:1 ratio of original to perturbed examples, using AdamW (8-bit) with learning rate 1×10^{-4}, cosine schedule, and a maximum sequence length of 16,384 tokens. Full hyperparameters are listed in Appendix B.1. For soft prompt tuning we set $m = 20$, while keeping all other hyperparameters the same as in the QLoRA setting.

4.3 Evaluation

We report accuracy per perturbation type (relative to the corresponding original examples) under each evaluation setting on the test split of the QuanTemp dataset explained in Section 3.1. We use strict scoring: only outputs that match the required JSON schema (`{"label": true/false}`) are included; all others are counted as parse failures and excluded (Appendix C.1). Parse failures are largely confined to the non-fine-tuned baseline and depend on prompting, whereas fine-tuned models exhibit near-zero failures. Thus, counting parse failures as incorrect would mainly lower baseline accuracy, with negligible effect on fine-tuned results.

	Model	Orig.	Approx.	Num.	Range	Neg.	Mask.	Rand.	M. Avg.
Base	$M_{GPT\text{-}5.4}$	78.12	92.70$^{+14.58}$	86.06$^{+7.94}$	93.44$^{+15.32}$	39.22$^{-38.90}$	17.00$^{-61.12}$	90.14$^{+12.02}$	73.96$^{-4.16}$
	$M_{Gemini\text{-}2.5}$	74.71	91.97$^{+17.26}$	86.05$^{+11.34}$	92.43$^{+17.72}$	53.85$^{-20.86}$	19.62$^{-55.09}$	92.06$^{+17.35}$	73.91$^{-0.80}$
	M_{Base}	86.35	85.05$^{-1.30}$	69.59$^{-16.76}$	82.35$^{-4.00}$	27.27$^{-59.08}$	8.87$^{-77.48}$	86.98$^{+0.63}$	62.80$^{-23.55}$
	$M_{Base\text{-}Two\text{-}Shot}$	81.96	88.72$^{+6.76}$	73.21$^{-8.75}$	84.78$^{+2.82}$	28.85$^{-53.11}$	12.44$^{-69.52}$	90.61$^{+8.65}$	66.10$^{-15.86}$
	$M_{32B\ Base}$	85.55	88.64$^{+3.09}$	78.89$^{-6.66}$	86.67$^{+1.12}$	36.00$^{-49.55}$	16.19$^{-69.36}$	86.41$^{+0.86}$	67.89$^{-17.66}$
	$M_{Only\text{-}Original}$	95.83	8.63$^{-87.20}$	7.41$^{-88.42}$	9.47$^{-86.36}$	7.69$^{-88.14}$	5.56$^{-90.27}$	12.50$^{-83.33}$	8.65$^{-87.18}$
All Perturb	$M_{PT\text{-}All\text{-}Perturbs}$	81.71	40.00$^{-41.71}$	32.06$^{-49.65}$	39.78$^{-41.93}$	26.53$^{-55.18}$	18.32$^{-63.39}$	52.38$^{-29.33}$	35.80$^{-45.91}$
	$M_{0.6B\ All\text{-}Perturbs}$	84.85	87.05$^{+2.20}$	76.85$^{-8.00}$	78.95$^{-5.90}$	**100.00**$^{+15.15}$	99.54$^{+14.69}$	95.37$^{+10.52}$	88.44$^{+3.59}$
	$M_{1.7B\ All\text{-}Perturbs}$	85.61	88.49$^{+2.88}$	78.24$^{-7.37}$	81.05$^{-4.56}$	98.08$^{+12.47}$	96.76$^{+11.15}$	93.52$^{+7.91}$	88.24$^{+2.63}$
	$M_{4B\ All\text{-}Perturbs}$	87.88	97.12$^{+9.24}$	95.37$^{+7.49}$	96.84$^{+8.96}$	**100.00**$^{+12.12}$	**100.00**$^{+12.12}$	97.69$^{+9.81}$	97.57$^{+9.69}$
	$M_{All\text{-}Perturbs}$	93.18	95.68$^{+2.50}$	96.30$^{+3.12}$	95.79$^{+2.61}$	98.08$^{+4.90}$	**100.00**$^{+6.82}$	**99.54**$^{+6.36}$	97.67$^{+4.49}$
Single Perturb	$M_{Only\text{-}Approx}$	91.67	96.40$^{+4.73}$	31.48$^{-60.19}$	83.16$^{-8.51}$	9.62$^{-82.05}$	28.24$^{-63.43}$	90.74$^{-0.93}$	60.45$^{-31.22}$
	$M_{Only\text{-}Num}$	95.45	93.53$^{-1.92}$	97.22$^{+1.77}$	95.79$^{+0.34}$	7.69$^{-87.76}$	21.76$^{-73.69}$	96.76$^{+1.31}$	76.00$^{-19.45}$
	$M_{Only\text{-}Range}$	93.94	67.63$^{-26.31}$	41.20$^{-52.74}$	97.37$^{+3.43}$	7.69$^{-86.25}$	10.19$^{-83.75}$	80.09$^{-13.85}$	55.10$^{-38.84}$
	$M_{Only\text{-}Negation}$	93.18	46.04$^{-47.14}$	29.17$^{-64.01}$	45.79$^{-47.39}$	98.08$^{+4.90}$	17.13$^{-76.05}$	59.26$^{-33.92}$	41.79$^{-51.39}$
	$M_{Only\text{-}Masked}$	93.56	24.46$^{-69.10}$	17.59$^{-75.97}$	26.32$^{-67.24}$	26.92$^{-66.64}$	**100.00**$^{+6.44}$	39.35$^{-54.21}$	42.47$^{-51.09}$
	$M_{Only\text{-}Random}$	94.70	95.68$^{+0.98}$	76.85$^{-17.85}$	93.68$^{-1.02}$	26.92$^{-67.78}$	58.80$^{-35.90}$	97.69$^{+2.99}$	80.56$^{-14.14}$
Exclusion	$M_{Excl\text{-}Original}$	94.70	97.84$^{+3.14}$	**98.61**$^{+3.91}$	**98.42**$^{+3.72}$	**100.00**$^{+5.30}$	99.54$^{+4.84}$	98.61$^{+3.91}$	**98.74**$^{+4.04}$
	$M_{Excl\text{-}Approx}$	91.67	97.84$^{+6.17}$	97.69$^{+6.02}$	96.84$^{+5.17}$	**100.00**$^{+8.33}$	**100.00**$^{+8.33}$	**99.54**$^{+7.87}$	98.54$^{+6.87}$
	$M_{Excl\text{-}Num}$	93.56	97.84$^{+4.28}$	62.04$^{-31.52}$	95.79$^{+2.23}$	**100.00**$^{+6.44}$	**100.00**$^{+6.44}$	98.61$^{+5.05}$	90.67$^{-2.89}$
	$M_{Excl\text{-}Range}$	**96.21**	95.68$^{-0.53}$	93.06$^{-3.15}$	95.26$^{-0.95}$	98.08$^{+1.87}$	99.54$^{+3.33}$	97.69$^{+1.48}$	96.41$^{+0.20}$
	$M_{Excl\text{-}Negation}$	91.67	**98.56**$^{+6.89}$	98.15$^{+6.48}$	96.84$^{+5.17}$	67.31$^{-24.36}$	**100.00**$^{+8.33}$	97.69$^{+6.02}$	96.70$^{+5.03}$
	$M_{Excl\text{-}Masked}$	93.56	95.68$^{+2.12}$	95.83$^{+2.27}$	96.84$^{+3.28}$	98.08$^{+4.52}$	13.43$^{-80.13}$	97.22$^{+3.66}$	79.10$^{-14.46}$
	$M_{Excl\text{-}Random}$	95.83	94.24$^{-1.59}$	94.91$^{-0.92}$	94.21$^{-1.62}$	98.08$^{+2.25}$	99.54$^{+3.71}$	93.98$^{-1.85}$	95.63$^{-0.20}$

Table 2: T→F accuracy (%) on unperturbed claims (**Orig.**, left of dotted rule) and label-flipping perturbations. Deltas vs. **Orig.** within each row. M. Avg. is the micro-average of accuracy across perturbation types.

All experiments use gold evidence to isolate the contribution of adversarial fine-tuning from retrieval quality. This controlled setup attributes observed performance differences unambiguously to the verification component rather than confounding them with retrieval errors; combining these gains with imperfect retrieval is an orthogonal question.

We report single-run results, as repeating inference across our >20 model configurations would be prohibitively expensive. The QuanTemp effects we discuss are large and consistent across perturbation types and evaluation directions, supporting confidence in the qualitative conclusions; for the cross-dataset, distance, and VitaminC comparisons, we test significance on the stored per-instance predictions (Sections 5.5–5.7).

5 Results

Tables 2–4 are organised into four model groups: Base Group (baselines and clean fine-tune), All Perturb Group (all-perturbation training), Single Perturb (single-type training), and Exclusion Group (leave-one-out training).

5.1 `True` → `False`

The T→F evaluation is the most diagnostic setting: numerical values are perturbed to flip the label, so a model that merely anchors on surface form will be confidently wrong.

The Base Group models' accuracy collapses for the *Negative* and *Masked* categories, with $M_{Only\text{-}Original}$ dropping by 83–90% on perturbed examples.

QLoRA models in All Perturb Group are far more robust than Base Group, reaching near-perfect accuracy on *Negative* and *Masked*. Unlike Single Perturb, which collapses on unseen types, All Perturb Group shows model size steadily improving robustness, with $M_{All\text{-}Perturbs}$ reaching mid-to-high-90s accuracy and the smallest fluctuations.

The leave-one-out group (Exclusion Group) has the best accuracies overall, but $M_{Excl\text{-}Num}$, $M_{Excl\text{-}Negation}$, and especially $M_{Excl\text{-}Masked}$ collapse on their omitted type (−31.52 pp, −24.36 pp, −80.13 pp). The most capable models are $M_{Excl\text{-}Original}$ and $M_{Excl\text{-}Range}$, with the smallest fluctuations across types. *Masked* detection does not transfer from any other training type; it must be

	Model	Orig.	Approx.	Num.	Range	Neg.	Mask.	Rand.	M. Avg.
Base	$M_{\text{GPT-5.4}}$	95.33	95.84$^{+0.51}$	96.51$^{+1.18}$	94.65$^{-0.68}$	92.22$^{-3.11}$	96.69$^{+1.36}$	97.75$^{+2.42}$	96.03$^{+0.70}$
	$M_{\text{Gemini-2.5}}$	91.48	90.77$^{-0.71}$	90.87$^{-0.61}$	89.61$^{-1.87}$	90.00$^{-1.48}$	88.55$^{-2.93}$	96.77$^{+5.29}$	91.35$^{-0.13}$
	M_{Base}	94.30	95.51$^{+1.21}$	94.65$^{+0.35}$	92.18$^{-2.12}$	93.75$^{-0.55}$	94.03$^{-0.27}$	97.70$^{+3.40}$	94.83$^{+0.53}$
	$M_{\text{Base-Two-Shot}}$	95.42	96.43$^{+1.01}$	95.99$^{+0.57}$	95.23$^{-0.19}$	94.51$^{-0.91}$	94.37$^{-1.05}$	98.60$^{+3.18}$	96.08$^{+0.66}$
	$M_{\text{32B Base}}$	95.24	96.64$^{+1.40}$	95.37$^{+0.13}$	94.67$^{-0.57}$	93.33$^{-1.91}$	95.77$^{+0.53}$	97.77$^{+2.53}$	95.96$^{+0.72}$
	$M_{\text{Only-Original}}$	96.75	97.70$^{+0.95}$	96.41$^{-0.34}$	96.92$^{+0.17}$	96.70$^{-0.05}$	96.61$^{-0.14}$	98.20$^{+1.45}$	97.13$^{+0.38}$
All Perturb	$M_{\text{PT-All-Perturbs}}$	94.65	96.84$^{+2.19}$	95.68$^{+1.03}$	97.31$^{+2.66}$	95.24$^{+0.59}$	95.69$^{+1.04}$	97.33$^{+2.68}$	96.35$^{+1.70}$
	$M_{\text{0.6B All-Perturbs}}$	95.12	94.90$^{-0.22}$	94.41$^{-0.71}$	84.60$^{-10.52}$	**100.00**$^{+4.88}$	**100.00**$^{+4.88}$	99.80$^{+4.68}$	95.27$^{+0.15}$
	$M_{\text{1.7B All-Perturbs}}$	96.10	94.64$^{-1.46}$	95.01$^{-1.09}$	92.18$^{-3.92}$	**100.00**$^{+3.90}$	99.80$^{+3.70}$	99.40$^{+3.30}$	96.55$^{+0.45}$
	$M_{\text{4B All-Perturbs}}$	**98.37**	97.45$^{-0.92}$	**97.80**$^{-0.57}$	93.60$^{-4.77}$	**100.00**$^{+1.63}$	**100.00**$^{+1.63}$	**100.00**$^{+1.63}$	98.01$^{-0.36}$
	$M_{\text{All-Perturbs}}$	97.40	96.94$^{-0.46}$	97.21$^{-0.19}$	95.02$^{-2.38}$	**100.00**$^{+2.60}$	**100.00**$^{+2.60}$	**100.00**$^{+2.60}$	98.05$^{+0.65}$
Single Perturb	$M_{\text{Only-Approx}}$	96.42	95.15$^{-1.27}$	96.61$^{+0.19}$	96.45$^{+0.03}$	95.60$^{-0.82}$	97.01$^{+0.59}$	99.00$^{+2.58}$	96.89$^{+0.47}$
	$M_{\text{Only-Num}}$	96.26	98.21$^{+1.95}$	96.61$^{+0.35}$	98.82$^{+2.56}$	94.51$^{-1.75}$	96.21$^{-0.05}$	**100.00**$^{+3.74}$	97.80$^{+1.54}$
	$M_{\text{Only-Range}}$	96.91	96.94$^{+0.03}$	96.61$^{-0.30}$	94.08$^{-2.83}$	96.70$^{-0.21}$	95.61$^{-1.30}$	99.20$^{+2.29}$	96.55$^{-0.36}$
	$M_{\text{Only-Negation}}$	95.12	96.94$^{+1.82}$	95.41$^{+0.29}$	96.68$^{+1.56}$	**100.00**$^{+4.88}$	95.81$^{+0.69}$	98.80$^{+3.68}$	96.84$^{+1.72}$
	$M_{\text{Only-Masked}}$	96.91	97.45$^{+0.54}$	96.61$^{-0.30}$	96.45$^{-0.46}$	96.70$^{-0.21}$	**100.00**$^{+3.09}$	98.60$^{+1.69}$	97.84$^{+0.93}$
	$M_{\text{Only-Random}}$	94.96	97.96$^{+3.00}$	95.41$^{+0.45}$	**99.53**$^{+4.57}$	97.80$^{+2.84}$	98.00$^{+3.04}$	**100.00**$^{+5.04}$	98.13$^{+3.17}$
Exclusion	$M_{\text{Excl-Original}}$	95.45	94.39$^{-1.06}$	96.01$^{+0.56}$	92.18$^{-3.27}$	**100.00**$^{+4.55}$	**100.00**$^{+4.55}$	99.80$^{+4.35}$	96.84$^{+1.39}$
	$M_{\text{Excl-Approx}}$	97.07	**99.23**$^{+2.16}$	97.01$^{-0.06}$	93.13$^{-3.94}$	**100.00**$^{+2.93}$	**100.00**$^{+2.93}$	**100.00**$^{+2.93}$	98.05$^{+0.98}$
	$M_{\text{Excl-Num}}$	96.75	96.43$^{-0.32}$	96.21$^{-0.54}$	91.47$^{-5.28}$	**100.00**$^{+3.25}$	**100.00**$^{+3.25}$	**100.00**$^{+3.25}$	97.14$^{+0.39}$
	$M_{\text{Excl-Range}}$	95.12	95.15$^{+0.03}$	94.61$^{-0.51}$	95.50$^{+0.38}$	**100.00**$^{+4.88}$	99.80$^{+4.68}$	99.60$^{+4.48}$	97.18$^{+2.06}$
	$M_{\text{Excl-Negation}}$	97.72	97.70$^{-0.02}$	**97.80**$^{+0.08}$	97.39$^{-0.33}$	97.80$^{+0.08}$	**100.00**$^{+2.28}$	**100.00**$^{+2.28}$	**98.63**$^{+0.91}$
	$M_{\text{Excl-Masked}}$	96.10	95.92$^{-0.18}$	96.01$^{-0.09}$	92.65$^{-3.45}$	**100.00**$^{+3.90}$	94.81$^{-1.29}$	**100.00**$^{+3.90}$	96.14$^{+0.04}$
	$M_{\text{Excl-Random}}$	94.63	94.64$^{+0.01}$	94.61$^{-0.02}$	91.94$^{-2.69}$	**100.00**$^{+5.37}$	**100.00**$^{+5.37}$	99.80$^{+5.17}$	96.55$^{+1.92}$

Table 3: F→F accuracy (%) on unperturbed claims (**Orig.**) and `False`-labelled perturbations. Deltas vs. **Orig.** within each row; M. Avg. is the micro-average of accuracy across perturbation types.

trained directly. Excluding *Numeration* hurts more than zero-shot, while excluding *Approximation*, *Range*, or *Random* has little negative effect (see Section 6).

5.2 `True` → `True`

Table 4 shows results for T→T, excluding *Negative*, *Random*, and *Masked* as they produce only non-structure-preserving perturbations.

The Base Group's and the All Perturb Group's models generally show the smallest accuracy fluctuations between original and perturbed examples. The $M_{\text{Only-Original}}$ model achieves mid-90s accuracy ($\pm 0.57\%$).

The Single Perturb and Exclusion Group models fluctuate more: single-type training provides no immunity to other perturbation types (e.g. $M_{\text{Only-Num}}$ drops -61.24% on *Range*), and leave-one-out training exposes similar gaps (e.g. $M_{\text{Excl-Range}}$ on *Range*, $M_{\text{Excl-Approx}}$ on *Approximation*). Adversarial fine-tuning is decisively beneficial: even $M_{\text{0.6B All-Perturbs}}$ exceeds $M_{\text{32B Base}}$ across all perturbation types, and excluding *Random* from training ($M_{\text{Excl-Random}}$) produces the strongest overall profile. The full analysis is in Section 6.

5.3 `False` → `False`

In the F→F evaluation (Table 3), models are robustly accurate, in the mid-90s across types. Fine-tuning on *Negative*, *Masked*, and *Random* saturates the small Base Group blind spots on those types. $M_{\text{Only-Random}}$ is nearly perfect (99.53%) on *Range*, where most other models struggle most. All models above 0.6B are robustly accurate on false claims; the main exceptions are *Range*, which causes accuracy drops across groups, and *Approximation*, where leave-one-out training counterintuitively yields the best accuracy (99.23%).

5.4 Cross-Dataset Generalisation

Table 5 shows how M_{Base}, $M_{\text{Excl-Approx}}$, and $M_{\text{All-Perturbs}}$ transfer to two unseen datasets. We restrict cross-dataset evaluation to baseline and all-perturbs models; other configurations generalised comparably and are omitted for brevity.

	Model	Approx.	Num.	Range	M. Avg.
Base	$M_{GPT\text{-}5.4}$	77.04$^{-1.08}$	81.43$^{+3.31}$	82.39$^{+4.27}$	80.28$^{+2.16}$
	$M_{Gemini\text{-}2.5}$	70.80$^{-3.91}$	74.65$^{-0.06}$	74.33$^{-0.38}$	73.56$^{-1.15}$
	M_{Base}	80.51$^{-5.84}$	87.44$^{+1.09}$	85.29$^{-1.06}$	84.92$^{-1.43}$
	$M_{Base\text{-}Two\text{-}Shot}$	73.33$^{-8.63}$	82.94$^{+0.98}$	79.56$^{-2.40}$	79.31$^{-2.65}$
	$M_{32B\ Base}$	77.44$^{-8.11}$	82.30$^{-3.25}$	81.01$^{-4.54}$	80.61$^{-4.94}$
	$M_{Only\text{-}Original}$	96.40$^{+0.57}$	**95.83**	95.26$^{-0.57}$	95.78$^{-0.05}$
All Perturb	$M_{PT\text{-}All\text{-}Perturbs}$	75.00$^{-6.71}$	80.10$^{-1.61}$	78.69$^{-3.02}$	77.93$^{-3.78}$
	$M_{0.6B\ All\text{-}Perturbs}$	89.21$^{+4.36}$	82.87$^{-1.98}$	88.95$^{+4.10}$	86.61$^{+1.76}$
	$M_{1.7B\ All\text{-}Perturbs}$	89.93$^{+4.32}$	88.89$^{+3.28}$	91.58$^{+5.97}$	90.09$^{+4.48}$
	$M_{4B\ All\text{-}Perturbs}$	89.93$^{+2.05}$	89.35$^{+1.47}$	91.05$^{+3.17}$	90.09$^{+2.21}$
	$M_{All\text{-}Perturbs}$	89.21$^{-3.97}$	91.20$^{-1.98}$	93.16$^{-0.02}$	91.38$^{-1.80}$
Single Perturb	$M_{Only\text{-}Approx}$	93.53$^{+1.86}$	93.98$^{+2.31}$	82.11$^{-9.56}$	89.73$^{-1.94}$
	$M_{Only\text{-}Num}$	71.94$^{-23.51}$	94.91$^{-0.54}$	34.21$^{-61.24}$	67.89$^{-27.56}$
	$M_{Only\text{-}Range}$	95.68$^{+1.74}$	93.98$^{+0.04}$	95.26$^{+1.32}$	94.86$^{+0.92}$
	$M_{Only\text{-}Negation}$	88.49$^{-4.69}$	92.59$^{-0.59}$	81.58$^{-11.60}$	87.71$^{-5.47}$
	$M_{Only\text{-}Masked}$	94.96$^{+1.40}$	94.44$^{+0.88}$	90.53$^{-3.03}$	93.21$^{-0.35}$
	$M_{Only\text{-}Random}$	58.27$^{-36.43}$	94.91$^{+0.21}$	18.42$^{-76.28}$	58.90$^{-35.80}$
Exclusion	$M_{Excl\text{-}Original}$	92.81$^{-1.89}$	94.44$^{-0.26}$	95.79$^{+1.09}$	94.49$^{-0.21}$
	$M_{Excl\text{-}Approx}$	53.96$^{-37.71}$	89.81$^{-1.86}$	93.16$^{+1.49}$	81.83$^{-9.84}$
	$M_{Excl\text{-}Num}$	95.68$^{+2.12}$	93.06$^{-0.50}$	95.26$^{+1.70}$	94.50$^{+0.94}$
	$M_{Excl\text{-}Range}$	96.40$^{+0.19}$	**95.83**$^{-0.38}$	65.79$^{-30.42}$	85.50$^{-10.71}$
	$M_{Excl\text{-}Negation}$	90.65$^{-1.02}$	91.20$^{-0.47}$	91.58$^{-0.09}$	91.19$^{-0.48}$
	$M_{Excl\text{-}Masked}$	93.53$^{-0.03}$	93.06$^{-0.50}$	95.79$^{+2.23}$	94.13$^{+0.57}$
	$M_{Excl\text{-}Random}$	**97.84**$^{+2.01}$	**95.83**	**96.32**$^{+0.49}$	**96.51**$^{+0.68}$

Table 4: T→T accuracy (%) on label-preserving perturbations of originally True claims (Negation, Random, Masked are False-only and excluded). Deltas vs. the unperturbed original; M. Avg. is the micro-average of accuracy across perturbation types.

Model	T1	T2	T3	M. Avg
AVeriTeC 2.0 dev				
M_{Base}	**95.65**	95.09	96.83	**95.31**
$M_{Excl\text{-}Approx}$	89.72$^{-5.93}$	93.73$^{-1.37}$	**97.20**$^{+0.37}$	92.13$^{-3.18}$
$M_{All\text{-}Perturbs}$	90.12$^{-5.53}$	**95.20**$^{+0.11}$	**97.20**$^{+0.37}$	93.20$^{-2.11}$
CLEF 2025 Spanish				
M_{Base}	**95.51**	97.17	97.40	96.44
$M_{Excl\text{-}Approx}$	94.83$^{-0.68}$	**99.64**$^{+2.47}$	**99.54**$^{+2.14}$	**97.48**$^{+1.04}$
$M_{All\text{-}Perturbs}$	94.71$^{-0.80}$	99.62$^{+2.45}$	99.44$^{+2.04}$	97.36$^{+0.92}$

Table 5: Cross-dataset generalisation accuracy (%). **T1** = value-preserving; **T2** = value-changing; **T3** = random/obfuscation. Deltas vs. M_{Base}. M. Avg. is the micro-averages of accuracy of **T1**, **T2** and **T3**.

CLEF 2025 Spanish. Both fine-tuned models beat M_{Base} overall (97.48% and 97.36% vs. 96.44%), with the largest gains on value-changing (T2) and obfuscation (T3) perturbations (99.44–99.54%, +2 pp). That aggregate understates the effect considerably. The split is 95.8% False, and the transition groups pool the instances whose label was flipped with the far larger set that was already False: the value-changing (T2) column combines 553 T→F instances with 5,578 F→F ones, so it tracks a majority on which every model is already near ceiling. Measured *within* transitions, $M_{All\text{-}Perturbs}$ improves T→F accuracy from 73.71% to 97.07% (+23.36 pp), a gain an order of magnitude larger than the aggregate suggests, and every other perturbation type improves by 13–25 pp. The largest gain by far is on *Negative* (+64.71 pp), where M_{Base} reaches only 35.29% – below chance – against 100.00% for both fine-tuned models. The gain is confined to the flipped cases: T→T is essentially unchanged (79.04% to 79.55%) and F→F is already near ceiling (97.44% to 98.28%), so fine-tuning does not make the model better at recognising claims that stay true, it stops it confirming claims whose quantity was altered. Appendix E gives the full per-type breakdown, and notes that M_{Base} leaves 75.6% of T→F instances unparseable, so its 73.71% rests on 175 of 717 instances.

AVeriTeC 2.0 dev. On parseable responses the baseline reaches 95.31% against 93.20% for $M_{All\text{-}Perturbs}$ (−2.11 pp), a deficit confined to T1; the two are within 0.4 pp on T2 and T3 (Table 5). It is not significant once paired ($p = 0.511$, Section 5.5) and rests on very different denominators: M_{Base} returns no parseable verdict on 281 of 750 instances (37.5%), against zero for the fine-tuned models. On AVeriTeC the advantage is coverage, not accuracy.

5.5 Statistical Significance

The comparisons so far are single-run point estimates. We recompute the cross-dataset results from the stored per-instance predictions with a paired bootstrap (10,000 resamples) and an exact McNemar test. Because M_{Base} leaves many instances unparseable while the fine-tuned models leave none, we report three protocols (Appendix D). On CLEF Spanish the gain is significant under every one: paired over the 9,090 instances both models parse, $M_{All\text{-}Perturbs}$ gains +2.33 pp ($p = 8.0e{-}41$) in every perturbation direction. Since that split is 95.8% False, macro-F1 is more informative: it rises 8.43 pp, almost entirely in the minority True class, whose F1 rises 16.4 pp from a baseline precision of just 42.4% (Table 15). On AVeriTeC the apparent parseable-only deficit is *not* significant once paired (−1.07 pp, $p = 0.511$); the real advantage there is coverage, as M_{Base} returns no parseable verdict on 37.5% of instances against zero for $M_{All\text{-}Perturbs}$. We therefore describe AVeriTeC as a coverage rather than an accuracy gain.

5.6 Sensitivity to Numerical Distance

Our taxonomy only coarsely distinguishes a perturbation that nudges a value from one that replaces it with an implausible figure. Binning each per-

turbed instance by relative numerical distance separates the two (Appendix F). The benefit of adversarial fine-tuning is largest exactly on the near-misses: +7.50 pp within 10% relative distance and +8.44 pp between 10% and 25%, falling monotonically to +2.91 pp beyond 100%, every bin individually significant. This bears directly on the concern that the models might be recognising perturbation artefacts rather than comparing quantities: an artefact detector would improve uniformly, or most where the edit is most conspicuous, whereas the observed profile is the opposite.

5.7 Evidence-Side Numerical Revisions

The perturbations studied so far all edit the *claim* while holding the evidence fixed, leaving open whether that robustness is direction-specific. We therefore evaluate on VitaminC (Schuster et al., 2021), which pairs each claim with two snapshots of one Wikipedia sentence, before and after a factual revision: the claim is unaltered natural text, and the quantity deciding the verdict changes in the *evidence*. Restricted to real, genuinely numerical revisions, it yields 11,887 test instances, including 5,227 *contrastive pairs* differing solely by a number in the evidence (Appendix G.1); we therefore also report *pair accuracy*, the share of pairs with *both* snapshots correct.

Claim-side robustness does not transfer. $M_{\text{All-Perturbs}}$ reaches only 78.35% accuracy and 54.45% pair accuracy, *below* the untuned M_{Base} at 90.96% and 84.25% (−12.60 pp and −29.81 pp, exact McNemar $p < 10^{-190}$; Table 18). The errors are strongly asymmetric: $M_{\text{All-Perturbs}}$ wrongly confirms a refuting snapshot on 31.5% of pairs but wrongly refutes a supporting one on only 15.0%, whereas M_{Base} is balanced (8.4% and 8.8%). Inspecting the 1,426 pairs where only $M_{\text{All-Perturbs}}$ errs shows a consistent pattern: "$24.9 million" is judged to support "grossed more than $25 million". The model checks whether the two quantities are approximately *equal* rather than evaluating the inequality the claim asserts: sufficient on our claim-side perturbations, which move the value away from the evidence, but not on bounded claims. Prompting does not repair it: Structured Chain-of-Thought (SCoT) and NCP are neutral or harmful, and a Threshold Comparison Prompt (TCP) naming the failure directly recovers only about half the deficit (Appendix G.4).

Training on evidence-side revisions closes it, and the two competencies are complementary. Trained with the same recipe on VitaminC itself, M_{VitaminC} reaches 98.05% accuracy and 96.56% pair accuracy, exceeding M_{Base} by +7.09 pp and +12.30 pp ($p < 10^{-140}$), and wrong confirmations fall from 31.5% to 2.5%. The recipe transfers even though the perturbation direction, the discrepancy's source and the evidence length all differ; the resulting *model* does not. Specialisation is costly in both directions (Table 6): M_{VitaminC} scores only 80.86% on QuanTemp against $M_{\text{All-Perturbs}}$'s 97.52%, below even M_{Base}. Training on both corpora jointly incurs no trade-off: M_{Mixed} attains 98.25% accuracy on VitaminC and 98.20% on QuanTemp, and is significantly better than *each specialist on that specialist's own test set* ($p = 1.0\text{e}{-4}$ and $p = 0.002$). Claim-side and evidence-side numerical robustness are thus distinct competencies, neither obtainable from the other, but jointly learnable.

6 Discussion

RQ1: Does adversarial fine-tuning improve robustness to label-flipping perturbations over clean-data fine-tuning? Yes: $M_{\text{Only-Original}}$ is near-perfect on value-preserving evaluations yet collapses to 5–13% across every label-flipping perturbation, an 83–90% drop. Standard fine-tuning thus teaches the model to confirm the original claim's surface form rather than verify it against evidence: without adversarial exposure, it never sees a plausible-looking claim that is numerically inconsistent with the evidence. All QLoRA models in All Perturb Group overcome this, matching or exceeding their T→T performance on every T→F perturbation.

F→F results are largely unaffected: false claims are already in the majority, and there is no truth-confirming prior to override. In T→T, all-perturbs variants lose 2.6–11.0% on *Original* relative to $M_{\text{Only-Original}}$, likely because the adversarial objective shifts the decision boundary away from truth-confirming features; Exclusion Group largely mitigates this, as excluding original data removes the tension.

RQ2: Does training on a perturbation subset generalise, and how does training composition affect robustness? In T→F, every Single Perturb model achieves ≥96% on its own perturbation type but collapses on at least one other. This is expected: each type exploits a distinct surface signal (digit-to-word rewriting, value-order shift, explicit negation),

Training data	VitaminC		QuanTemp		QuanTemp by perturbation type				
	Acc.	Pair	Acc.	M-F1	Orig.	T1	T2	Mask	Rand
None (M_{Base})	90.96	84.25	—	—	—	—	—	—	—
QuanTemp ($M_{All\text{-}Perturbs}$)	78.35	54.45	97.52	94.62	96.13	94.95	98.80	**100.00**	**99.86**
VitaminC ($M_{VitaminC}$)	98.05	96.56	80.86	69.80	79.29	72.69	92.26	57.88	95.12
Both (M_{Mixed})	**98.25**	**97.15**	**98.20**	**96.10**	**96.36**	**96.45**	**99.45**	**100.00**	99.72

Table 6: Each training regime on both test sets (VitaminC: 11,887 instances, 5,227 pairs; QuanTemp: 6,175). Specialising on either corpus alone costs 17–20 pp on the other, so neither competency is a by-product of the other; training on both recovers it and beats each specialist on its own test set (+0.68 pp QuanTemp, $p = 1.0\text{e}{-4}$; +0.59 pp pair accuracy VitaminC, $p = 0.002$). **T1** = value-preserving; **T2** = value-changing; **Mask/Rand** are the obfuscation variants.

so training on one gives no immunity to the others. The partial exception is $M_{Only\text{-}Random}$, which transfers positively across F→F: replacing any number with an arbitrary value trains a general "distrust the number" reflex rather than a specific pattern.

The Exclusion Group ablation shows that not all perturbation types are equally interchangeable. Excluding *Masked* or *Numeration* degrades T→F performance below even the zero-shot baseline: *Masked* introduces a unique token pattern (###) that no other type replicates, and *Numeration* requires learning digit-to-word equivalences that cannot be inferred from other rewriting patterns. By contrast, excluding *Approximation* or *Random* has a negligible effect: these signals are recoverable from cues already present in the other types.

Training composition also interacts with evaluation direction: excluding *Approximation* yields the best F→F accuracy (99.23%) but collapses in T→T (53.96%), as the model no longer recognises "about X" as truth-preserving. Excluding *Random* produces the strongest T→T profile with little effect elsewhere. Finally, $M_{Excl\text{-}Original}$, trained only on perturbed data, recovers almost all T→T accuracy lost by All Perturb Group, confirming that unperturbed training examples, not the perturbations themselves, shift the decision boundary.

RQ3: How do model scale and adaptation method interact with adversarial fine-tuning? Scale matters, but fine-tuning matters more: $M_{0.6B\ All\text{-}Perturbs}$ beats zero-shot $M_{32B\ Base}$, GPT-5.4 Pro and Gemini 2.5 Flash on T→F by at least 14 pp in micro-average accuracy. Across the all-perturbs series, T→F accuracy is flat at 88% for 0.6B and 1.7B, jumps to 98% at 4B, and saturates at 8B; the jump comes mostly from *Numeration*, *Range*, and *Approximation*. Both 4B and 8B outperform every Base Group model by at least 23 pp.

QLoRA models in All Perturb Group and Exclusion Group consistently outperform soft-prompt tuning. Soft prompts can shift the output distribution but do not update the weight-level representations needed to detect numerical inconsistencies; QLoRA modifies attention and feed-forward projections directly. Soft-prompt tuning still beats $M_{Only\text{-}Original}$ on T→F, so adversarial exposure helps, but it stays below the zero-shot baseline, suggesting the prepended vectors introduce conflicting biases rather than targeted numerical sensitivity.

RQ4: Does adversarial fine-tuning generalise across domains and languages? Across languages, yes; on AVeriTeC, the gain is in coverage rather than accuracy (Sections 5.4–5.5). On CLEF 2025 Spanish the evidence is unstructured article text, as in QuanTemp, so the shift is primarily lexical and the structural signal transfers across languages. On AVeriTeC the baseline retains a narrow edge on label-flipping, where all examples are False: its tendency to predict False inflates the score, the artefact already seen in QuanTemp T→F.

7 Conclusion

This paper is a diagnostic study of LLM robustness to numerical claim perturbations. Standard fine-tuning teaches the model to confirm a claim's surface form rather than verify it against evidence, and accuracy on label-flipping inputs collapses to single digits. Adversarial fine-tuning on counterfactual examples recovers near-perfect accuracy, with small QLoRA-tuned models (0.6B–8B) surpassing zero-shot 32B and frontier closed-source baselines on the hardest perturbations. Composition matters more than variety: *Masked* and *Numeration* are essential, *Approximation* and *Random* redundant. Robustness also transfers across languages and domains without target-language data, suggesting a generalised numerical decision boundary rather than memorised edits.

Limitations

Data and perturbation scope. Our perturbation taxonomy rewrites numerical expressions in the claim while leaving evidence unchanged. Section 5.7 reaches the evidence side through VitaminC's human-authored revisions, but our generator does not produce them, and jointly perturbing claim and evidence remains an open direction. The six types focus on scalar quantities as digits or written-out numbers, so robustness to unit conversions, date arithmetic, compound quantities, or numbers in tables is not guaranteed. Training and primary evaluation use QuanTemp (English fact-checking); cross-dataset experiments add Spanish, AVeriTeC and VitaminC but remain in English and Spanish and, apart from VitaminC's single-sentence evidence, within the fact-checking domain, so broader language coverage and generalisation to scientific, financial, or tabular settings both require further evaluation.

Model and evaluation setup. Experiments use only the Qwen3 family (0.6B–32B) with QLoRA and soft prompt tuning. Aarnes and Setty (2025) document the same numerical brittleness on LLaMA-family models, so we expect the qualitative findings to hold across LLaMA, Mistral, Gemma, and comparable open families, though confirming this directly is left for future work; quantitative effects (the scale at which diminishing returns set in, the size of the QLoRA-vs.-soft-prompt gap) may shift across families and PEFT methods. The soft-prompt baseline reuses QLoRA hyperparameters, making it a matched-budget rather than optimally tuned comparison. All experiments assume gold evidence, isolating verification from retrieval; whether the robustness gains survive noisy or incomplete retrieval is left for future work.

Interpretability. We lack a mechanistic account of *why* adversarial fine-tuning works. It is unclear whether models learn to attend more carefully to numerical tokens, develop surface-form-invariant representations, or rely on other signals; attribution, probing, and activation-patching analyses would be needed to explain the gains and anticipate failure modes — a direction we intend to pursue.

Ethical Considerations

This work studies the robustness of LLMs used for numerical claim verification. We identify specific weaknesses that existing models exhibit when claims contain perturbed numerical expressions, and we show that adversarial fine-tuning substantially reduces these vulnerabilities. We do not endorse any specific model or configuration for deployment in real-world fact-checking pipelines; fact-checking is a nuanced task where model outputs should be treated as decision-support tools rather than authoritative verdicts, and human oversight remains essential.

The perturbation techniques studied could, in principle, be used to craft adversarial claims that evade automated fact-checkers. We consider this a reason to publish rather than withhold the findings: understanding and documenting these failure modes is a prerequisite for building more resilient systems, and the perturbation types we study are straightforward enough that they are already available to any motivated adversary. All data used in this work is sourced from publicly available fact-checking corpora, and no personal or sensitive information is involved.

Acknowledgements

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

References

Peter Røysland Aarnes and Vinay Setty. 2025. NumPert: Numerical perturbations to probe language models for veracity prediction. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 78–95.

Mubashara Akhtar, Rami Aly, Yulong Chen, Zhenyun Deng, Michael Schlichtkrull, Chenxi Whitehouse, and Andreas Vlachos. 2025. The 2nd automated verification of textual claims (AVeriTeC) shared task: Open-weights, reproducible and efficient systems. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 201–223.

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*, pages 15391–15405.

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian

Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and V. Venktesh. 2025. Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 199–223.

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*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2025. MathOdyssey: Benchmarking mathematical problem-solving skills in large language models using Odyssey math data. *Scientific Data*, 12(1):1392.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *10th International Conference on Learning Representations*.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *8th International Conference on Learning Representations*.

Malin Astrid Larsson, Harald Fosen Grunnaleite, and Vinay Setty. 2026. One LLM to train them all: Multitask learning framework for fact-checking. In *Advances in Information Retrieval – 48th European Conference on Information Retrieval*, pages 289–304.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*.

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 Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10647–10660.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*, pages 65128–65167.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024. NumeroLogic: Number encoding for enhanced LLMs' numerical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 206–212.

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*, pages 2842–2846.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, pages 8818–8847.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

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*, pages 650–660.

V Venktesh, Vinay Setty, Avishek Anand, Boushra Bendou, Maram Hasanain, Houda Bouamor, Gabriel Iturra-Bocaz, Petra Galuščáková, and Firoj Alam. 2025. Overview of the CLEF-2025 CheckThat! Lab task 3 on fact-checking numerical claims. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 709–721.

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.

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*, pages 7950–7966.

Yuli Yang, Hiroaki Yamada, and Takenobu Tokunaga. 2025. Evaluating robustness of LLMs to numerical variations in mathematical reasoning. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 171–180.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2024a. 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.

Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. 2024b. MathAttack: Attacking large language models towards math solving ability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19750–19758.

A Datasets and Perturbations

A.1 QuanTemp Split Statistics

Tables 7 and 8 present the sizes of the training and validation datasets, respectively. In the training split, evidence documents contain a median of 535 words per sample; the validation split has a median of 521 words and a mean of 604; the test split (see Table 1) has a median of 535 words and a mean of 648.

The maximum word count across the data splits is 6,943.

Category	T→T	T→F	F→F	Total
Numeration	775	775	3872	5422
Approximation	539	539	3010	4088
Range	931	931	4868	6730
Masked	✗	775	1939	2714
Random	✗	775	1939	2714
Negative	✗	218	294	512

Table 7: QuanTemp **train** split: sample counts by perturbation type and transition direction. Original claims: 931 `True`, 2,434 `False`.

Category	T→T	T→F	F→F	Total
Numeration	273	273	1206	1752
Approximation	197	197	938	1332
Range	324	324	1494	2142
Masked	✗	273	603	876
Random	✗	273	603	876
Negative	✗	75	95	170

Table 8: QuanTemp **validation** split: sample counts by perturbation type and transition direction. Original claims: 324 `True`, 747 `False`.

A.2 Perturbation Type Definitions and Examples

Table 9 defines each perturbation type with a concrete example. Types marked T1 generate both a value-preserving (T1) and a value-changing (T2) variant; types marked T2 or T3 generate only that variant.

A.3 Cross-Dataset Evaluation Sets

AVeriTeC 2.0 dev (Schlichtkrull et al., 2023; Akhtar et al., 2025) is an English claim verification dataset drawn from real-world fact-checking sources. We filter its development split[4] to numerical claims, retaining 119 unique claim–evidence pairs (49 `True`, 70 `False`). Evidence documents are relatively short (average 114 words), formatted as retrieved question–answer pairs rather than

[4]Ground-truth labels for the test split are not publicly released.

Type	Label	Original claim fragment	Perturbed fragment	Variant
Numeration	T1	"fell to **3.5%**"	"fell to **three point five percent**" "fell to **thirty-five percent**"	T1 (preserving) T2 (changing)
Approximation	T1	"matched **50 years**"	"matched **about 50 years**" "matched **about 5 years**"	T1 (preserving) T2 (changing)
Range	T1	"over **30 million** filed"	"over **25–35 million** filed" "over **10–15 million** filed"	T1 (preserving) T2 (changing)
Masked	T3	"rate of **3.5%**"	"rate of **####**"	T3 only
Random	T3	"fell by **12 points**"	"fell by **47 points**"	T3 only
Negation	T2	"quarterly earnings was **5%**"	"quarterly earnings was **-5%**"	T2 only

Table 9: Perturbation type definitions and examples (Aarnes and Setty, 2025). T1 = value-preserving; T2 = value-changing; T3 = random replacement and masked obfuscation.

	AVeriTeC 2.0 dev		CLEF 2025 Spanish	
Category	**True**	**False**	**True**	**False**
Num.	74	124	328	3,614
Approx.	64	90	322	3,470
Range	64	94	317	3,397
Rand.	45	62	164	1,807
Neg.	5	9	70	338
Total	252	379	1,201	12,626

Table 10: Evaluated perturbed instances for the cross-dataset generalisation splits, by label of the source claim. Numeration, Approximation and Range each contribute a value-preserving (T1) and a value-changing (T2) variant per claim, both of which remain `False` for `False` source claims; only the *Random* obfuscation variant is evaluated, and perturbations whose text duplicates another instance are removed. **AVeriTeC 2.0 dev**: 119 claims (49T / 70F), 750 instances including originals; **CLEF 2025 Spanish**: 2,633 claims (200T / 2,433F), 16,460 instances including originals.

free-form passages. We use an LLM (Claude Opus 4.7) to draft a free-form evidence passage based on the claim, question and answer texts. We apply five perturbation categories (*Numeration*, *Approximation*, *Range*, *Random*, *Negative*), yielding 631 perturbed instances after removing perturbations whose text duplicates another instance (Table 10).

CLEF 2025 Spanish is derived from a Spanish-language fact-checking corpus from the CLEF CheckThat! 2025 challenge (Venktesh et al., 2025). After filtering to numerically grounded claims containing at least one stand-alone quantity (excluding identifier-embedded numbers such as *5G* or *COVID-19*), we retain 2,633 unique claim–evidence pairs. The split is heavily skewed towards `False` claims (200 `True` vs. 2,433 `False`), reflecting AFP-style debunking corpora that primarily publish corrections to false claims. Evidence documents are substantially longer (median 584 words, mean 642 words) as they include full fact-check articles. Applying the same five perturbation types and deduplication yields 13,827 perturbed instances (16,460 including the originals), making it the largest evaluation set in this study.

B Training and Inference Details

B.1 Training Hyperparameters

Hyperparameters used for Hugging Face's SFTTrainer[5] configuration can be found in Table 11. Additional SFTTrainer parameters are set to their default values.

Hyperparameter	Value
Epochs	2
Batch size (per device)	5
Gradient accumulation	1
Learning rate	1×10^{-4}
LR scheduler	Cosine
Warmup steps	50
Optimiser	AdamW (8-bit)
Weight decay	0.01
Max sequence length	16,384
Precision	bfloat16 / fp16
Quantisation	4-bit NF4
LoRA rank r	32
LoRA α	16
LoRA dropout	0
LoRA target modules	q, k, v, o, gate, up, down proj

Table 11: QLoRA training hyperparameters.

B.2 Soft Prompt Tuning Formulation

A sequence of m continuous prompt vectors $P = [p_1, \ldots, p_m] \in \mathbb{R}^{m \times d}$ is prepended to the frozen input embeddings $E(x_c, x_e)$, giving the initial hidden state:

$$h^{(0)} = \left[P\,;\, E(x_c, x_e)\right]$$

Only P and the veracity head $\{w, b\}$ are updated; Θ is fully frozen. Veracity is predicted as $\hat{y} = \sigma(w^\top h^{(L)}_{\text{cls}} + b)$, where $h^{(L)}_{\text{cls}}$ is the final-layer representation of the prefix position. The number of

[5] https://huggingface.co/docs/trl/sft_trainer

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.

Instruction prompt

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

{Two-shot or NCP examples}

Claim: `{claim}`

Evidence: `{evidence}`

Figure 3: Prompt template for binary fact-checking. The model receives a claim and an evidence passage, then returns a JSON label indicating whether the claim is true or false.

Snippet of NCP Prompt

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.

Reasoning: The claim states that the Eiffel Tower is three hundred and fifty-one meters tall (351 m), but the evidence indicates it is 330 meters tall. Since the numbers do not match, the claim is false.

{"label": false}

Figure 4: Approximation NCP example. The full prompt provides one example per perturbation type and is added to the instruction prompt.

trainable parameters is $m \cdot d$, independent of model size.

With $m = 20$, soft prompts are particularly weak on label-flipping perturbations, which require rewriting (not just nudging) the decision boundary. Because this gap was already large in pilot runs, we did not pursue further soft-prompt hyperparameter tuning: closing it would require orders of magnitude more capacity, which would not address the structural limitation that soft prompts cannot modify the verification computation.

B.3 Prompts

Figure 3 displays the template used for the system and instruction (zero-shot) prompt used during fine-tuning and evaluation – where {claim} and {evidence} signal the input data. In the two-shot prompt evaluation, the model is provided with an additional sentence informing it that it will receive two samples before its actual task – all being added in the {Two-shot sample position}. Full two-shot examples can be accessed through our GitHub repository[6].

Numerical Counterexample Prompt (*NCP*) is further examined for the Qwen3-8B base models (q4_k_m). The prompt includes the zero-shot instructions and additional short, trivial counterexamples illustrating why a numerical claim with each perturbation type is false.

[6] https://github.com/iai-group/numpert-fine-tune

B.4 Infrastructure and Hardware

Computational resources are provided by a small GPU cluster managed with the Slurm Workload Manager. We do not report GPU-hours, as fine-tuning and inference jobs are scheduled dynamically across the available hardware rather than assigned to a fixed GPU type. The cluster is heterogeneous, comprising a small number of NVIDIA H100 80GB, NVIDIA A100 40GB, and Tesla V100 GPUs.

C Output Parsing

C.1 Parse-Failure Rates

Table 12 presents the parse-failure macro averages for the main results. Parse failures are mostly localised to non-fine-tuned models in Base Group, varying substantially across prompting settings in the main evaluation. M_{Base} produces non-trivial error rates in the Zero-Shot setting (up to 17.90% on T→F), which Two-Shot prompting nearly eliminates (below 3% across directions). $M_{\text{32B Base}}$ exhibits consistently low error rates ($\leq 4.40\%$) across all settings, suggesting that parsing failures are largely a function of model scale rather than prompting strategy. GPT-5.4 Pro and Gemini 2.5 Flash parsing errors are 2.5–4% and 6.6–16.2%, respectively. The unquantised counterparts exhibit higher failure rates – M_{Base} between 40.49% and 56.25% and $M_{\text{32B Base}}$ between 9.90% and 25.78% – though Two-Shot prompting improves this, reducing failures to roughly 10–15%.

Model	T→F	T→T	F→F
Gemini	14.38	16.19	6.64
GPT-5.4	3.00	4.00	2.50
Quantised (Q4_K_M)			
M_{Base}	17.90	10.30	11.90
$M_{Base\text{-}Two\text{-}Shot}$	2.40	2.80	0.50
$M_{32B\ Base}$	4.40	3.70	1.30
$M_{PT\text{-}All\text{-}Perturbs}$	9.70	0.00	0.00
Unquantised			
M_{Base}	56.25	40.49	55.54
$M_{Base\text{-}Two\text{-}Shot}$	14.93	10.14	12.39
$M_{32B\ Base}$	25.78	16.16	9.90
$M_{Only\text{-}Masked}$	6.62	4.00	4.34
$M_{Only\text{-}Random}$	–	8.25	10.20

Table 12: Parse failure rates (%) by split and quantisation.

C.2 NCP Under Relaxed Parsing

The high NCP parse-failure rate is largely a strict-parser artefact. Of the 3,599 NCP outputs rejected by the strict JSON parser, 96.8% contain a valid `{"label": true/false}` substring embedded in a longer response (typically followed by a free-form reasoning paragraph). Under a relaxed parser that extracts the first such substring anywhere in the response, the headline NCP parse-failure rate drops from 58.3% to 1.9% (115 truly unparseable outputs out of 6,175), and overall NCP accuracy rises from 85.68% to 90.17%.

Table 13 compares the strict and relaxed $M_{Base\text{-}NCP}$ rows across all three evaluation directions. The recovery is not uniform. On label-flipping (T→F) and label-preserving False (F→F) settings the recovered cells improve substantially (the most striking change is Negation in T→F, $47.83 \rightarrow 72.55\%$, where strict parsing was selectively rejecting the model's correct flipping decisions). On label-preserving True (T→T) settings the recovered cells worsen ($79.05 \rightarrow 71.50\%$ macro average), indicating that on these inputs the model's verbose-output mode produces parseable-but-incorrect predictions which strict parsing was inadvertently filtering out. We retain strict-parser numbers in the main tables as a consistent protocol across all configurations and report the relaxed numbers here for transparency.

Column	Strict	Relaxed	Δ
T→F			
Orig.	76.16	73.93	−2.23
Approx.	89.66	94.93	+5.27
Num.	81.00	85.92	+4.92
Range	87.23	91.94	+4.71
Neg.	47.83	72.55	+24.72
Mask.	30.53	33.15	+2.62
Rand.	86.46	92.49	+6.03
M. Avg.	72.20	78.50	+6.30
F→F			
Orig.	91.74	94.79	+3.05
Approx.	94.53	95.90	+1.37
Num.	93.68	95.98	+2.30
Range	77.27	87.32	+10.05
Neg.	94.74	96.70	+1.96
Mask.	89.95	92.37	+2.42
Rand.	98.82	99.00	+0.18
M. Avg.	91.28	94.55	+3.27
T→T			
Approx.	64.63	59.56	−5.07
Num.	83.94	79.81	−4.13
Range	84.03	75.14	−8.89
M. Avg.	79.05	71.50	−7.55

Table 13: $M_{Base\text{-}NCP}$ accuracy (%) under strict (used in main tables) vs. relaxed JSON parsing. Relaxed parsing extracts the first `{"label":...}` substring anywhere in the response, recovering 96.8% of strict-parser rejections (overall parse-failure rate drops from 58.3% to 1.9%). Recovery improves T→F and F→F cells but worsens T→T cells.

D Statistical Significance

We recompute the cross-dataset results from the stored per-instance predictions with a paired bootstrap (10,000 resamples) and an exact McNemar test (Table 14). Because M_{Base} leaves a substantial share of instances unparseable while the fine-tuned models leave none, the treatment of those instances is itself a design choice, so we report three protocols: excluding them, as in Table 5; restricting to the instances both models parse; and scoring them as errors.

CLEF Spanish. The gain is significant under every protocol. On the 9,090 instances both models parse, $M_{All\text{-}Perturbs}$ improves overall accuracy by +2.33 pp (CI $[+1.99, +2.69]$, $p = 8.0\text{e}{-41}$) and wins in every perturbation direction: +1.96 pp on Original ($p = 2.1\text{e}{-6}$), +2.30 pp on T1 ($p = 1.2\text{e}{-12}$), +2.63 pp on T2 ($p = 1.4\text{e}{-21}$) and +2.21 pp on T3 ($p = 2.4\text{e}{-7}$). Since the split is 95.8% `False`, accuracy sits near the majority-class rate and macro-F1 is the more informative statistic: it rises from 76.88 to 85.31 (+8.43 pp, CI $[+6.26, +10.72]$). Table 15 shows the improvement is almost entirely in the minority `True` class,

Dataset	Protocol	M_{Base}	$M_{All\text{-}Perturbs}$	Δ (pp)	95% CI (pp)	McNemar p
CLEF Spanish	Parseable-only (reported)	96.44	97.36	+0.93	[+0.53, +1.35]	—
	Paired (n = 9,090)	96.44	98.77	+2.33	[+1.99, +2.69]	8.0e−41
	Failures as errors	53.26	97.36	+44.11	[+43.34, +44.86]	$< 10^{-300}$
AVeriTeC	Parseable-only (reported)	95.31	93.20	−2.11	[−4.51, +0.33]	—
	Paired (n = 469)	95.31	94.24	−1.07	[−3.60, +1.46]	0.511
	Failures as errors	59.60	93.20	+33.60	[+29.87, +37.33]	4.4e−57

Table 14: Significance of the cross-dataset gains (overall accuracy, %). Deltas are $M_{All\text{-}Perturbs}$ minus M_{Base}, with a 10,000-resample paired bootstrap and an exact McNemar test on stored per-instance predictions. The three protocols differ only in how M_{Base}'s unparseable responses are handled: excluded (as in Table 5), restricted to the instances *both* models parse, or scored as errors. On CLEF Spanish the gain is significant under every protocol; on AVeriTeC the apparent parseable-only deficit is not significant once the comparison is paired ($p = 0.511$), and the fine-tuned model's real advantage there is coverage: M_{Base} returns no parseable verdict on 281 of 750 instances (37.5%) against zero for $M_{All\text{-}Perturbs}$.

	True			False			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**M-F1**
M_{Base}	42.4	80.9	55.6	**99.4**	96.9	98.1	76.88
$M_{Excl\text{-}Approx}$	**65.7**	**82.5**	**73.1**	99.2	**98.1**	**98.7**	**85.89**
$M_{All\text{-}Perturbs}$	64.4	81.6	72.0	99.2	98.0	98.6	85.31

Table 15: Per-class results on CLEF 2025 Spanish (%). The split is 95.8% `False`, so overall accuracy is close to the majority-class rate and macro-F1 (**M-F1**) is the informative figure: fine-tuning improves it by 8.4–9.0 pp, and the improvement is concentrated almost entirely in the minority `True` class (+16.4 pp F1 for $M_{All\text{-}Perturbs}$), where M_{Base}'s precision is only 42.4%. Support differs across rows because M_{Base} is scored on the 9,090 instances it parses (251 `True`) whereas the fine-tuned models are scored on all 16,460 (684 `True`); M_{Base} fails to parse 7,370 instances (44.8%) against zero for either fine-tuned model.

whose F1 rises by 16.4 pp from a baseline precision of just 42.4%; the `False` class was already near ceiling for both models.

AVeriTeC. Here the parseable-only comparison suggests $M_{All\text{-}Perturbs}$ is 2.11 pp *worse* than M_{Base}, but that difference is not significant once the comparison is paired over the 469 instances both models answer (−1.07 pp, CI [−3.60, +1.46], p = 0.511). The fine-tuned model's actual advantage on this dataset is coverage: M_{Base} produces no parseable verdict on 281 of 750 instances (37.5%), against zero for $M_{All\text{-}Perturbs}$, so under the failures-as-errors protocol the ordering reverses to +33.60 pp (p = 4.4e−57). We therefore describe the AVeriTeC result as a coverage gain with indistinguishable accuracy where both models commit, rather than as an accuracy gain.

E CLEF Spanish: Per-Perturbation Breakdown

Table 5 reports CLEF Spanish only at the level of the transition groups, which aggregate over the label transition. On a split that is 95.8% `False` this conceals the effect of interest, because each group pools the instances whose label was flipped with the far larger set whose label was already `False`: the value-changing (T2) column combines 553 T→F instances with 5,578 F→F ones, and the T3 column 164 with 1,807. An aggregate therefore tracks the F→F majority, where every model is already near ceiling, rather than the label-flipping cases the perturbations were designed to probe.

Table 16 reports accuracy per perturbation type *within* each transition. Three observations follow.

The label-flipping gain is an order of magnitude larger than the aggregate suggests. On T→F, $M_{All\text{-}Perturbs}$ improves from 73.71% to 97.07% (+23.36 pp) and $M_{Excl\text{-}Approx}$ to 97.49% (+23.78 pp), against the +2.45 pp visible in the aggregated T2 column. Every perturbation type improves: *Approximation* by 25.16 pp, *Random* by 20.25 pp, *Numeration* by 16.34 pp and *Range* by 13.70 pp.

***Negative* is the baseline's worst failure and the largest single gain.** M_{Base} reaches only 35.29% on negated Spanish claims, below chance for a binary decision, while both fine-tuned models reach 100.00%. Negation inverts a percentage or rate, so the surface form stays close to the evidence, and a model that matches on surface overlap rather than comparing the quantity will confirm it; this mirrors the same type being the hardest T→F case for the frontier models on QuanTemp.

The gain is confined to the flipped cases. On T→T the fine-tuned model is statistically indistinguishable from the baseline (79.04% to 79.55%), and on F→F both are near ceiling (97.44% to 98.28%). Adversarial fine-tuning on Spanish therefore does not make the model better at recognising claims that remain true; it stops the model from confirming claims whose quantity has been altered. This is the same asymmetry visible in the per-class results of Table 15, where the `True`-class gain is driven by precision rather than recall.

Model	Approx.	Num.	Range	Neg.	Rand.	M. Avg.
T→F *(label-flipping)*						
M_{Base}	72.97	80.00	82.50	35.29	76.09	73.71
$M_{Excl\text{-}Approx}$	$98.76^{+25.78}$	$97.56^{+17.56}$	$95.57^{+13.07}$	$100.00^{+64.71}$	$96.95^{+20.86}$	$97.49^{+23.78}$
$M_{All\text{-}Perturbs}$	$98.14^{+25.16}$	$96.34^{+16.34}$	$96.20^{+13.70}$	$100.00^{+64.71}$	$96.34^{+20.25}$	$97.07^{+23.36}$
(n)	*161*	*164*	*158*	*70*	*164*	*717*
T→T *(label-preserving)*						
M_{Base}	78.43	82.26	75.93	–	–	79.04
$M_{Excl\text{-}Approx}$	$64.60^{-13.84}$	$85.98^{+3.72}$	$89.31^{+13.38}$	–	–	$79.96^{+0.92}$
$M_{All\text{-}Perturbs}$	$77.64^{-0.79}$	$79.27^{-2.99}$	$81.76^{+5.84}$	–	–	$79.55^{+0.50}$
(n)	*161*	*164*	*159*	–	–	*484*
F→F *(already false)*						
M_{Base}	97.85	97.55	96.31	97.28	98.39	97.44
$M_{Excl\text{-}Approx}$	$99.14^{+1.29}$	$98.34^{+0.79}$	$96.53^{+0.22}$	$100.00^{+2.72}$	$99.78^{+1.39}$	$98.32^{+0.88}$
$M_{All\text{-}Perturbs}$	$98.33^{+0.48}$	$98.51^{+0.96}$	$97.06^{+0.75}$	$100.00^{+2.72}$	$99.72^{+1.33}$	$98.28^{+0.84}$
(n)	*3470*	*3614*	*3397*	*338*	*1807*	*12626*

Table 16: CLEF 2025 Spanish accuracy (%) by perturbation type *within* each label transition. Deltas are against M_{Base} in the same column. Accuracy is computed on each model's parseable responses, as in Table 5. **Neg.** and **Rand.** produce `False` labels by construction and so cannot appear under T→T. Reporting within transitions separates the 717 label-flipping instances from the 15,059 F→F ones that dominate this split; the aggregated columns of Table 5 pool them and understate the label-flipping gain by an order of magnitude.

Parse failures are concentrated in the hard cases. The accuracies above follow Table 5 in scoring each model on its parseable responses, which flatters M_{Base} unevenly: it fails to parse 75.6% of T→F instances, 63.3% of T→T and 42.5% of F→F, against zero for either fine-tuned model. Its 73.71% on T→F therefore rests on 175 of 717 instances, and on a subset it selected by answering at all. Scoring failures as errors instead, M_{Base} falls to 17.99% on T→F, 29.68% on T→T and 56.01% on F→F, so the T→F gap widens from +23.36 pp to +79.08 pp. The true value lies between the two protocols, but under either the label-flipping gain is far larger than the aggregated table indicates.

Relative distance	n	M_{Base}	$M_{All\text{-}Perturbs}$	Δ
0 (value preserved)	1,373	90.17	96.07	+5.90
$(0, 10\%]$	399	87.99	95.49	+7.50
$(10\%, 25\%]$	345	90.11	98.55	+8.44
$(25\%, 50\%]$	661	94.14	99.39	+5.26
$(50\%, 100\%]$	513	94.06	98.83	+4.77
$> 100\%$	503	96.69	99.60	+2.91

Table 17: Accuracy (%) by the relative distance between the original and the perturbed quantity, on the QuanTemp test split (3,794 of 4,579 perturbed instances align to a distance, 83% coverage). M_{Base} is scored on parseable responses only. Every bin is individually significant (exact McNemar, $p < 10^{-2}$ throughout and $p < 10^{-7}$ in three bins). Near-miss values are the harder case for both models, and the gain from adversarial fine-tuning is largest exactly there (+7.5 to +8.4 pp below 25% distance) and smallest for implausible replacements beyond 100%, which is the opposite of what a model that had merely learned to recognise perturbation artefacts would show.

F Sensitivity to Numerical Distance

The taxonomy of Section 3.1 encodes numerical distance only coarsely. We therefore bin each perturbed instance by the relative distance between the original and the perturbed quantity, scoring range perturbations as 0 when the value falls inside the interval and otherwise as the distance to the nearest bound. Of 4,579 perturbed test instances, 3,794 (83%) align to a distance; Table 17 reports accuracy per bin.

Near-miss values are the harder case for both models, and the benefit of adversarial fine-tuning is largest exactly there: +7.50 pp within 10% relative distance and +8.44 pp between 10% and 25%, falling monotonically to +2.91 pp beyond 100%. Every bin is individually significant. This bears directly on the concern that the fine-tuned models might be recognising perturbation artefacts rather than comparing quantities: an artefact detector would improve uniformly across the distance axis, or most where the edit is most conspicuous, whereas the observed profile is the opposite and

tracks the difficulty of the numerical comparison itself.

G VitaminC: Evidence-Side Revisions

G.1 Subset Construction and Splits

VitaminC contains 488,904 claim–evidence pairs, of which 163,180 (33.4%) are annotator-written synthetic revisions rather than real Wikipedia edits. We retain only real revisions, and only the `SUPPORTS`/`REFUTES` labels that map onto our binary protocol, discarding `NOT ENOUGH INFO`. We further require that the revision is genuinely numerical, i.e. that the multiset of numeric tokens differs between the two evidence snapshots, and that the claim carries a quantitative entity (CARDINAL, MONEY, PERCENT or QUANTITY), applying the same spaCy tagging used to build our QuanTemp splits. Approximately 47% of VitaminC's real revisions alter a numeric token by this measure, consistent with the 48% quantitative rate reported from manual annotation (Schuster et al., 2021), which is what makes the corpus a dense source of numerical contrasts.

Splits. We build the subset with the filters above applied to each official VitaminC split. The test split yields 11,887 instances over 3,701 revisions (6,650 `True`, 5,237 `False`). Because the same claim recurs against both snapshots, these instances include 5,227 *contrastive pairs*: two evaluations of one claim whose gold labels differ solely because a number in the evidence changed. The dev split yields 16,434 instances over 5,143 revisions. The splits are disjoint not only by revision and claim text but by Wikipedia page, since VitaminC is partitioned at the page level, so no article contributes to more than one split. We use dev only for prompt selection, and never for model selection or reporting.

Evidence length. VitaminC evidence is a single revised sentence, with a median of 26 words (mean 28). For comparison, our QuanTemp splits have a median of 535 words (Appendix A.1) and AVeriTeC a mean of 114 words (Appendix A.3). The corpus therefore isolates numerical comparison from long-context retrieval, and it constitutes a distribution shift for models trained on QuanTemp's much longer passages – a shift that works *against* the fine-tuned models, making their deficit in Table 18 conservative rather than inflated.

G.2 Pair Accuracy

Instance accuracy understates how hard this setting is, because a model that ignores the quantity and keys off claim–evidence surface overlap can still score well on the supporting snapshot of every pair. We therefore report *pair accuracy*: the share of contrastive pairs for which *both* snapshots are labelled correctly. This is the direct analogue of our label-flipping evaluation, and we additionally decompose its errors into the two directions – wrongly refuting a supporting snapshot, and wrongly confirming a refuting one – because their asymmetry identifies the failure mode rather than merely its size.

G.3 Training Details

Training sample. From the training split we draw 16,000 instances by sampling whole revisions rather than instances, so that a contrastive pair is never divided; revisions carrying both labels are preferred, and in the resulting sample all 4,044 revisions are contrastive and the label distribution is balanced (7,997 `True` / 8,003 `False`). Validation is a further 2,001 instances over 508 revisions, drawn from the training split and verified case-disjoint from it, which leaves the official dev split untouched. The sample is sized to be comparable to the 17,258 QuanTemp training records rather than to exhaust the 77,568 available VitaminC instances.

Deduplication. Our pipeline deduplicates training samples by claim text, which is correct for QuanTemp but destructive here: because VitaminC pairs one claim with two evidence documents, keying on text alone discards one snapshot of every pair, reducing the 16,000-instance sample to 7,987 and removing precisely the contrast the corpus provides. We therefore key deduplication on the (claim text, gold label) pair, which retains 15,952 of 16,000 VitaminC samples (99.7%) while changing the QuanTemp training set by only 234 samples (≈1%). Two samples sharing a claim but carrying opposite gold labels are contrastive, not duplicate.

Optimisation. M_{VitaminC} and M_{Mixed} are trained from the base model with the hyperparameters of Appendix B.1 (rank 32, batch size 5, no gradient accumulation, two epochs). M_{VitaminC} uses a maximum sequence length of 2,048, since VitaminC's longest prompt is 648 tokens. M_{Mixed} is trained on the union of the QuanTemp all-perturbations recipe and the VitaminC sample, 34,567 samples

Model	Prompt	Per instance Acc.	Per instance Macro-F1	Per pair Pair Acc.	Pair error mode Refutes	Pair error mode Confirms	PF
No target-domain training							
M_{Base}	Zero-shot	90.96	90.86	84.25	8.8	8.4	704
Trained on QuanTemp claim-side perturbations							
$M_{All\text{-}Perturbs}$	Zero-shot	$78.35^{-12.60}$	77.63	$54.45^{-29.81}$	15.0	31.5	0
$M_{All\text{-}Perturbs}$	SCoT	$74.98^{-15.98}$	73.20	$46.15^{-38.10}$	10.7	44.2	0
$M_{All\text{-}Perturbs}$	NCP	$76.18^{-14.78}$	76.00	$53.74^{-30.51}$	24.2	23.2	0
$M_{All\text{-}Perturbs}$	TCP	$85.16^{-5.80}$	84.85	$69.16^{-15.09}$	12.1	19.5	0
Trained with VitaminC evidence-side revisions							
$M_{VitaminC}$	Zero-shot	$98.05^{+7.09}$	98.02	$96.56^{+12.30}$	1.2	2.5	0
M_{Mixed}	Zero-shot	$\mathbf{98.25}^{+7.29}$	**98.23**	$\mathbf{97.15}^{+12.89}$	1.4	1.7	0

Table 18: Results on the VitaminC numerical test subset (11,887 instances; 5,227 contrastive pairs). **Pair Acc.** is the share of contrastive pairs for which both snapshots of a revision are labelled correctly. **Refutes** is the share of pairs whose supporting snapshot is wrongly refuted, and **Confirms** the share whose refuting snapshot is wrongly confirmed; the asymmetry between them identifies the failure mode. **PF** counts parse failures, which are scored as errors throughout so that all rows share the same denominator. Deltas are against M_{Base}. SCoT = Structured Chain-of-Thought, NCP = Numerical Counterexample Prompt, TCP = Threshold Comparison Prompt.

in total, at a maximum sequence length of 8,192: the mixed corpus has a median prompt length of 450 tokens and a 99th percentile of 3,521, and only 0.03% of samples exceed 8,192. For the mixed run we leave the perturbation-ratio, T→F ratio and method-balancing constraints unset, as they are defined over QuanTemp's perturbed-versus-original structure and would otherwise treat the VitaminC instances as material against which to balance perturbations.

G.4 Additional Results

Prompting does not close the gap. Because the deficit looks like a decision rule rather than missing knowledge, we ask whether prompting can recover it. Neither of our existing strategies helps: SCoT, which explicitly instructs the model to extract, normalise and compare quantities step by step, *lowers* pair accuracy to 46.15% and raises wrong confirmations to 44.2%, and NCP is approximately neutral at 53.74%. We additionally designed a Threshold Comparison Prompt (TCP) that names the failure directly, stating that a claim asserting a bound holds only if the evidence value satisfies that comparison and not merely if the two numbers are close, illustrated with invented figures so that no test instance enters the prompt. Selected on the VitaminC dev split (+5.09 pp accuracy, CI $[+4.56, +5.60]$; +11.08 pp pair accuracy, CI $[+9.99, +12.17]$), it transfers to the test split with +6.81 pp accuracy (CI $[+6.18, +7.42]$) and +14.71 pp pair accuracy (CI $[+13.45, +15.97]$), both $p < 10^{-100}$. Even so, TCP recovers only about half the deficit and remains 15.09 pp below M_{Base} in pair accuracy. Notably, under every prompt the fine-tuned model emits a bare verdict of roughly 15 characters and never produces the reasoning SCoT demands: adversarial fine-tuning appears to have fixed the output format at the cost of instruction-following, which is why prompts can shift its verdicts but cannot restore deliberation.

Head-to-head deltas. The deficit of $M_{All\text{-}Perturbs}$ against M_{Base} is -12.60 pp accuracy (95% CI $[-13.42, -11.77]$) and -29.81 pp pair accuracy (CI $[-31.36, -28.24]$), exact McNemar $p < 10^{-190}$. Against $M_{All\text{-}Perturbs}$, $M_{VitaminC}$ gains +19.69 pp accuracy and +42.11 pp pair accuracy on VitaminC. M_{Mixed} is significantly better than each specialist on that specialist's own test set: +0.68 pp over $M_{All\text{-}Perturbs}$ on QuanTemp (CI $[+0.34, +1.02]$, $p = 1.0\text{e}{-4}$, with 78 instances that only M_{Mixed} answers correctly against 36 for $M_{All\text{-}Perturbs}$) and +0.59 pp pair accuracy over $M_{VitaminC}$ on VitaminC (CI $[+0.23, +0.96]$, $p = 0.002$).

Validity checks. Three checks guard the comparison. First, we re-ran $M_{All\text{-}Perturbs}$ on the QuanTemp test split through the same code path used for all VitaminC evaluations and reproduced the published figure to within three instances of 6,175 (97.47% against 97.52%), which excludes a pipeline regression as the explanation for its VitaminC result. Second, the VitaminC splits share no revision, no claim text and no Wikipedia page, so the evaluation is clean by construction and the TCP prompt was selected on dev rather than test. Third, all models are evaluated at the same 4-bit quantisation and

with parse failures scored as errors, which is conservative for M_{Base}: it is the only model with a non-trivial failure rate (704 of 11,887, 5.92%), and on parseable responses alone it reaches 96.68%. We note as a scope condition that M_{Base} produces several hundred tokens of free-form reasoning per instance, whereas the fine-tuned models emit a single verdict, so part of the baseline's strength on this shorter-evidence corpus reflects inference-time computation that the fine-tuned models do not spend.

H Unquantised Model Results

Tables 19, 20 and 21 report the same evaluation as Section 5 but use unquantised 16-bit checkpoints (BF16, or FP16 on GPUs without BF16 support) rather than the q4_k_m quantisation used in the main paper. Models are organised into the same four groups as in the main results: Base Group (baselines and clean fine-tune), All Perturb Group (all-perturbation training), Single Perturb (single-type training), and Exclusion Group (leave-one-out training). The qualitative patterns mirror the quantised results closely across all three evaluation directions: clean fine-tuning collapses on label-flipping perturbations, all-perturbs training restores near-perfect accuracy, single-type training fails to generalise across perturbation types, and the same perturbation types remain essential (*Masked*, *Numeration*) or redundant (*Approximation*, *Random*). Per-cell numbers differ by at most a few percentage points in either direction, but the relative ordering of model groups and the qualitative conclusions of the main paper hold.

H.1 `True` → `False`

Base Group collapses on *Negative* and *Masked*, the absolute outlier being $M_{Only\text{-}Original}$ with 94.70% on originals but only 9.42% micro-average (−85.28 pp) and 79–87% drops across every perturbed category – closely tracking the collapse seen in the quantised setting (Table 2). All Perturb Group is uniformly more robust, with $M_{All\text{-}Perturbs}$ showing the lowest accuracy fluctuations in its group. Single Perturb reaches perfect accuracy on its fine-tuned perturbation type but collapses elsewhere, again matching the single-type generalisation pattern reported in the main results. Exclusion Group achieves the best overall accuracies: $M_{Excl\text{-}Masked}$ drops −75.85% on its held-out perturbation type (close to the −80.13% reported for the quantised counterpart), while $M_{Excl\text{-}Approx}$ and $M_{Excl\text{-}Range}$ *improve* on their held-out type, confirming that the *Approximation* and *Range* signals are recoverable from compositional cues in the remaining training mix.

H.2 `True` → `True`

Table 20 presents the results for the T→T evaluation. In sharp contrast to T→F, Base Group dominates: M_{Base} reaches a micro-average of 96.69%, while All Perturb Group trails in the 83–91% range. This is the same direction observed in the quantised results (Table 4), where all-perturbs training loses ground on label-preserving inputs because the adversarial objective shifts the decision boundary away from truth-confirming features. Single Perturb and Exclusion Group instabilities concentrate on the *Range* perturbation type, again consistent with the quantised findings.

H.3 `False` → `False`

In the F→F evaluation (Table 21), most models achieve robust accuracies of 96–99%, indicating that fine-tuning largely resolves the small blind spots observed in Base Group– identical to the pattern in the quantised setting (Table 3). The leave-one-out models show little degradation on their held-out perturbation types, with $M_{Excl\text{-}Approx}$ and $M_{Excl\text{-}Range}$ both scoring near saturation.

Summary. Across all three evaluation directions, the unquantised results reproduce the qualitative findings of the q4_k_m experiments: the same model groups, perturbation types, and headline conclusion, that adversarial fine-tuning mitigates label-flipping brittleness at a small cost to label-preserving accuracy, hold independently of quantisation.

	Model	Orig.	Approx.	Num.	Range	Neg.	Mask.	Rand.	Micro Avg.
Base	M_{Base}	94.35	81.82$^{-12.53}$	35.94$^{-58.41}$	78.18$^{-16.17}$	3.70$^{-90.65}$	2.50$^{-91.85}$	71.43$^{-22.92}$	48.74$^{-45.61}$
	$M_{\text{Base-Two-Shot}}$	86.92	86.61$^{-0.31}$	62.50$^{-24.42}$	80.00$^{-6.92}$	21.74$^{-65.18}$	8.63$^{-78.29}$	84.70$^{-2.22}$	60.28$^{-26.64}$
	$M_{\text{32B Base}}$	88.24	85.86$^{-2.38}$	66.42$^{-21.82}$	85.34$^{-2.90}$	29.27$^{-58.97}$	7.29$^{-80.95}$	83.80$^{-4.44}$	61.90$^{-26.34}$
	$M_{\text{Only-Original}}$	**94.70**	8.63$^{-86.07}$	6.94$^{-87.76}$	9.47$^{-85.23}$	7.69$^{-87.01}$	6.94$^{-87.76}$	15.28$^{-79.42}$	9.42$^{-85.28}$
All Perturb	$M_{\text{0.6B All-Perturbs}}$	81.06	93.53$^{+12.47}$	75.00$^{-6.06}$	83.68$^{+2.62}$	**100.00**$^{+18.94}$	99.54$^{+18.48}$	96.76$^{+15.70}$	90.09$^{+9.03}$
	$M_{\text{1.7B All-Perturbs}}$	85.23	89.93$^{+4.70}$	88.43$^{+3.20}$	88.42$^{+3.19}$	98.08$^{+12.85}$	99.54$^{+14.31}$	94.91$^{+9.68}$	92.81$^{+7.58}$
	$M_{\text{4B All-Perturbs}}$	86.36	**100.00**$^{+13.64}$	95.37$^{+9.01}$	98.42$^{+12.06}$	**100.00**$^{+13.64}$	**100.00**$^{+13.64}$	98.61$^{+12.25}$	98.44$^{+12.08}$
	$M_{\text{All-Perturbs}}$	92.42	97.84$^{+5.42}$	96.76$^{+4.34}$	96.84$^{+4.42}$	**100.00**$^{+7.58}$	**100.00**$^{+7.58}$	99.54$^{+7.12}$	98.35$^{+5.93}$
Single Perturb	$M_{\text{Only-Approx}}$	92.05	96.40$^{+4.35}$	36.11$^{-55.94}$	85.79$^{-6.26}$	21.15$^{-70.90}$	36.57$^{-55.48}$	92.13$^{+0.08}$	64.53$^{-27.52}$
	$M_{\text{Only-Num}}$	92.42	92.81$^{+0.39}$	98.15$^{+5.73}$	96.84$^{+4.42}$	13.46$^{-78.96}$	28.24$^{-64.18}$	95.83$^{+3.41}$	77.75$^{-14.67}$
	$M_{\text{Only-Range}}$	92.05	72.66$^{-19.39}$	42.13$^{-49.92}$	97.37$^{+5.32}$	19.23$^{-72.82}$	11.11$^{-80.94}$	78.70$^{-13.35}$	56.46$^{-35.59}$
	$M_{\text{Only-Negation}}$	88.64	54.68$^{-33.96}$	36.57$^{-52.07}$	50.53$^{-38.11}$	98.08$^{+9.44}$	19.91$^{-68.73}$	68.98$^{-19.66}$	48.01$^{-40.63}$
	$M_{\text{Only-Masked}}$	90.20	29.01$^{-61.19}$	23.38$^{-66.82}$	33.15$^{-57.05}$	27.08$^{-63.12}$	99.50$^{+9.30}$	42.05$^{-48.15}$	46.03$^{-44.17}$
	$M_{\text{Only-Random}}$	92.24	97.62$^{+5.38}$	74.87$^{-17.37}$	94.61$^{+2.37}$	38.78$^{-53.46}$	59.30$^{-32.94}$	98.41$^{+6.17}$	81.44$^{-10.80}$
Exclusion	$M_{\text{Excl-Original}}$	93.56	98.56$^{+5.00}$	**99.07**$^{+5.51}$	98.42$^{+4.86}$	**100.00**$^{+6.44}$	**100.00**$^{+6.44}$	98.61$^{+5.05}$	**99.03**$^{+5.47}$
	$M_{\text{Excl-Approx}}$	88.64	99.28$^{+10.64}$	98.15$^{+9.51}$	97.37$^{+8.73}$	**100.00**$^{+11.36}$	**100.00**$^{+11.36}$	99.54$^{+10.90}$	98.93$^{+10.29}$
	$M_{\text{Excl-Num}}$	89.39	97.12$^{+7.73}$	67.59$^{-21.80}$	96.84$^{+7.45}$	**100.00**$^{+10.61}$	**100.00**$^{+10.61}$	98.61$^{+9.22}$	91.93$^{+2.54}$
	$M_{\text{Excl-Range}}$	92.80	94.96$^{+2.16}$	95.83$^{+3.03}$	95.79$^{+2.99}$	98.08$^{+5.28}$	99.54$^{+6.74}$	97.69$^{+4.89}$	96.99$^{+4.19}$
	$M_{\text{Excl-Negation}}$	89.02	99.28$^{+10.26}$	98.61$^{+9.59}$	97.37$^{+8.35}$	65.38$^{-23.64}$	**100.00**$^{+10.98}$	97.69$^{+8.67}$	96.89$^{+7.87}$
	$M_{\text{Excl-Masked}}$	92.05	95.68$^{+3.63}$	96.76$^{+4.71}$	97.89$^{+5.84}$	98.08$^{+6.03}$	16.20$^{-75.85}$	98.15$^{+6.10}$	80.27$^{-11.78}$
	$M_{\text{Excl-Random}}$	94.32	94.24$^{-0.08}$	95.83$^{+1.51}$	95.26$^{+0.94}$	98.08$^{+3.76}$	99.54$^{+5.22}$	95.83$^{+1.51}$	96.40$^{+2.08}$

Table 19: T→F accuracy (%) for Qwen3-8B (unquantized fp16), selected models. Columns as in Table 2.

	Model	Orig.	Approx.	Num.	Range	Micro Avg.
Base	M_{Base}	94.35	93.67$^{-0.68}$	97.33$^{+2.98}$	**98.17**$^{+3.82}$	**96.69**$^{+2.34}$
	$M_{\text{Base-Two-Shot}}$	86.92	80.83$^{-6.09}$	86.14$^{-0.78}$	89.41$^{+2.49}$	85.93$^{-0.99}$
	$M_{\text{32B Base}}$	88.24	89.29$^{+1.05}$	90.37$^{+2.13}$	87.84$^{-0.40}$	89.21$^{+0.97}$
	$M_{\text{Only-Original}}$	**94.70**	95.68$^{+0.98}$	95.37$^{+0.67}$	94.21$^{-0.49}$	95.04$^{+0.34}$
All Perturb	$M_{\text{0.6B All-Perturbs}}$	81.06	84.17$^{+3.11}$	78.70$^{-2.36}$	87.37$^{+6.31}$	83.12$^{+2.06}$
	$M_{\text{1.7B All-Perturbs}}$	85.23	89.93$^{+4.70}$	85.65$^{+0.42}$	91.58$^{+6.35}$	88.81$^{+3.58}$
	$M_{\text{4B All-Perturbs}}$	86.36	89.21$^{+2.85}$	88.89$^{+2.53}$	90.00$^{+3.64}$	89.36$^{+3.00}$
	$M_{\text{All-Perturbs}}$	92.42	87.77$^{-4.65}$	90.74$^{-1.68}$	92.63$^{+0.21}$	90.64$^{-1.78}$
Single Perturb	$M_{\text{Only-Approx}}$	92.05	91.37$^{-0.68}$	93.06$^{+1.01}$	74.74$^{-17.31}$	86.24$^{-5.81}$
	$M_{\text{Only-Num}}$	92.42	71.94$^{-20.48}$	93.52$^{+1.10}$	27.37$^{-65.05}$	64.95$^{-27.47}$
	$M_{\text{Only-Range}}$	92.05	92.81$^{+0.76}$	92.13$^{+0.08}$	93.68$^{+1.63}$	92.84$^{+0.79}$
	$M_{\text{Only-Negation}}$	88.64	84.89$^{-3.75}$	90.74$^{+2.10}$	76.32$^{-12.32}$	84.22$^{-4.42}$
	$M_{\text{Only-Masked}}$	90.20	89.63$^{-0.57}$	89.66$^{-0.54}$	87.43$^{-2.77}$	88.87$^{-1.33}$
	$M_{\text{Only-Random}}$	92.24	58.27$^{-33.97}$	91.54$^{-0.70}$	23.56$^{-68.68}$	59.36$^{-32.88}$
Exclusion	$M_{\text{Excl-Original}}$	93.56	94.24$^{+0.68}$	93.06$^{-0.50}$	94.74$^{+1.18}$	93.95$^{+0.39}$
	$M_{\text{Excl-Approx}}$	88.64	55.40$^{-33.24}$	87.04$^{-1.60}$	90.53$^{+1.89}$	80.19$^{-8.45}$
	$M_{\text{Excl-Num}}$	89.39	92.81$^{+3.42}$	87.50$^{-1.89}$	93.16$^{+3.77}$	90.83$^{+1.44}$
	$M_{\text{Excl-Range}}$	92.80	94.96$^{+2.16}$	93.98$^{+1.18}$	52.11$^{-40.69}$	79.63$^{-13.17}$
	$M_{\text{Excl-Negation}}$	89.02	92.09$^{+3.07}$	86.57$^{-2.45}$	90.53$^{+1.51}$	89.36$^{+0.34}$
	$M_{\text{Excl-Masked}}$	92.05	93.53$^{+1.48}$	92.59$^{+0.54}$	92.63$^{+0.58}$	92.84$^{+0.79}$
	$M_{\text{Excl-Random}}$	94.32	96.40$^{+2.08}$	92.13$^{-2.19}$	94.74$^{+0.42}$	94.13$^{-0.19}$

Table 20: T→T accuracy (%) for Qwen3-8B (unquantised, BF16/FP16), selected models. Columns as in Table 4.

	Model	Orig.	Approx.	Num.	Range	Neg.	Mask.	Rand.	Micro Avg.
Base	M_{Base}	91.69	95.73$^{+4.04}$	94.37$^{+2.68}$	86.88$^{-4.81}$	88.10$^{-3.59}$	92.52$^{+0.83}$	95.91$^{+4.22}$	92.98$^{+1.29}$
	$M_{\text{Base-Two-Shot}}$	94.95	96.77$^{+1.82}$	96.30$^{+1.35}$	92.31$^{-2.64}$	93.10$^{-1.85}$	93.62$^{-1.33}$	98.33$^{+3.38}$	95.42$^{+0.47}$
	$M_{\text{32B Base}}$	95.99	96.69$^{+0.70}$	95.87$^{-0.12}$	94.48$^{-1.51}$	93.75$^{-2.24}$	95.50$^{-0.49}$	98.21$^{+2.22}$	96.09$^{+0.10}$
	$M_{\text{Only-Original}}$	97.56	97.45$^{-0.11}$	97.21$^{-0.35}$	96.92$^{-0.64}$	95.60$^{-1.96}$	96.81$^{-0.75}$	98.40$^{+0.84}$	97.30$^{-0.26}$
All Perturb	$M_{\text{0.6B All-Perturbs}}$	95.93	95.15$^{-0.78}$	95.41$^{-0.52}$	88.86$^{-7.07}$	**100.00**$^{+4.07}$	**100.00**$^{+4.07}$	**100.00**$^{+4.07}$	96.30$^{+0.37}$
	$M_{\text{1.7B All-Perturbs}}$	96.59	94.64$^{-1.95}$	95.21$^{-1.38}$	92.18$^{-4.41}$	**100.00**$^{+3.41}$	**100.00**$^{+3.41}$	99.40$^{+2.81}$	96.64$^{+0.05}$
	$M_{\text{4B All-Perturbs}}$	**98.37**	98.21$^{-0.16}$	97.60$^{-0.77}$	94.79$^{-3.58}$	**100.00**$^{+1.63}$	**100.00**$^{+1.63}$	**100.00**$^{+1.63}$	98.30$^{-0.07}$
	$M_{\text{All-Perturbs}}$	98.21	96.43$^{-1.78}$	97.41$^{-0.80}$	94.08$^{-4.13}$	**100.00**$^{+1.79}$	**100.00**$^{+1.79}$	**100.00**$^{+1.79}$	97.84$^{-0.37}$
Single Perturb	$M_{\text{Only-Approx}}$	97.24	95.66$^{-1.58}$	97.01$^{-0.23}$	96.68$^{-0.56}$	95.60$^{-1.64}$	98.00$^{+0.76}$	99.40$^{+2.16}$	97.38$^{+0.14}$
	$M_{\text{Only-Num}}$	97.40	98.47$^{+1.07}$	97.21$^{-0.19}$	99.53$^{+2.13}$	95.60$^{-1.80}$	97.41$^{+0.01}$	**100.00**$^{+2.60}$	98.38$^{+0.98}$
	$M_{\text{Only-Range}}$	97.56	97.19$^{-0.37}$	97.60$^{+0.04}$	94.31$^{-3.25}$	97.80$^{+0.24}$	97.21$^{-0.35}$	99.60$^{+2.04}$	97.30$^{-0.26}$
	$M_{\text{Only-Negation}}$	96.42	97.70$^{+1.28}$	97.41$^{+0.99}$	98.10$^{+1.68}$	**100.00**$^{+3.58}$	96.81$^{+0.39}$	99.00$^{+2.58}$	97.88$^{+1.46}$
	$M_{\text{Only-Masked}}$	97.31	97.85$^{+0.54}$	97.50$^{+0.19}$	97.27$^{-0.04}$	95.40$^{-1.91}$	**100.00**$^{+2.69}$	99.16$^{+1.85}$	98.30$^{+0.99}$
	$M_{\text{Only-Random}}$	96.22	98.57$^{+2.35}$	97.22$^{+1.00}$	**100.00**$^{+3.78}$	98.73$^{+2.51}$	99.10$^{+2.88}$	**100.00**$^{+3.78}$	**98.95**$^{+2.73}$
Exclusion	$M_{\text{Excl-Original}}$	96.91	95.66$^{-1.25}$	97.01$^{+0.10}$	94.79$^{-2.12}$	**100.00**$^{+3.09}$	**100.00**$^{+3.09}$	**100.00**$^{+3.09}$	97.76$^{+0.85}$
	$M_{\text{Excl-Approx}}$	98.05	98.98$^{+0.93}$	97.80$^{-0.25}$	95.26$^{-2.79}$	**100.00**$^{+1.95}$	**100.00**$^{+1.95}$	**100.00**$^{+1.95}$	98.55$^{+0.50}$
	$M_{\text{Excl-Num}}$	98.21	97.45$^{-0.76}$	97.21$^{-1.00}$	93.84$^{-4.37}$	**100.00**$^{+1.79}$	**100.00**$^{+1.79}$	**100.00**$^{+1.79}$	97.92$^{-0.29}$
	$M_{\text{Excl-Range}}$	97.07	96.43$^{-0.64}$	96.41$^{-0.66}$	98.58$^{+1.51}$	**100.00**$^{+2.93}$	**100.00**$^{+2.93}$	**100.00**$^{+2.93}$	98.42$^{+1.35}$
	$M_{\text{Excl-Negation}}$	**98.37**	97.96$^{-0.41}$	98.40$^{+0.03}$	97.63$^{-0.74}$	97.80$^{-0.57}$	**100.00**$^{+1.63}$	**100.00**$^{+1.63}$	98.84$^{+0.47}$
	$M_{\text{Excl-Masked}}$	97.24	96.43$^{-0.81}$	97.60$^{+0.36}$	94.08$^{-3.16}$	**100.00**$^{+2.76}$	97.01$^{-0.23}$	**100.00**$^{+2.76}$	97.26$^{+0.02}$
	$M_{\text{Excl-Random}}$	96.91	95.92$^{-0.99}$	96.21$^{-0.70}$	94.08$^{-2.83}$	**100.00**$^{+3.09}$	**100.00**$^{+3.09}$	**100.00**$^{+3.09}$	97.51$^{+0.60}$

Table 21: F→F accuracy (%) for Qwen3-8B (unquantized fp16), selected models. Columns as in Table 3