Qeios

Peer Review

# Review of: "A Dutch Financial Large Language Model"

**Evgenii Arsentev**[1]

1. Independent researcher

## *What the paper claims*

FinGEITje is a 7B Dutch financial model built by QLoRA instruction tuning of GEITje-7B, itself a Dutch continuation of Mistral-7B-v0.1. The training set holds 147,788 Dutch instructions (Table 3), machine-translated from eight open English financial corpora with gpt-3.5-turbo-0125 at temperature 0. Alongside the model, the authors release the translation and filtering pipeline, the instruction data, and the first Dutch financial evaluation benchmark covering sentiment analysis, headline classification, named entity recognition, relation extraction, and question answering. A second LLM is used at evaluation time to map free-form generations onto the expected label set. On the Dutch benchmark, FinGEITje leads every task; on English data, it wins NER and places second on three of the remaining four.

The engineering is real, and the artefacts are public: model card, datasets, benchmark, and code are all linked and reachable. For anyone adapting an open model to a smaller language, that package is worth more than the leaderboard line. My objections below are about what the numbers can carry, not about the value of the release.

## *Does the conclusion hold on the data?*

Three things stand between the tables and the claim of across-the-board superiority.

**1. The paper never says that the benchmark items were held out from instruction tuning.** Table 2 (training) and Table 4 (evaluation) draw on the same corpora, and the deduplication in Section 3.1.4 is exact-match only. Exact matching cannot catch the fan-out the pipeline itself produces: in Table 2, the NER (CLS) row turns 609 raw samples into 17,051 instructions, and headlines go from 11,412 to 102,708, a ninefold expansion. Table 4 confirms that the NER task in the benchmark is that same NER (CLS) set. If the split was made after that expansion, the same source sentence sits on both sides wearing different

 doi.org/10.32388/G98QNJ 

instruction wrappers, and part of the reported margin is memorisation. One sentence stating where the split was cut, at the raw-sample level or after expansion, would settle the question; a near-duplicate check over the final split would settle it convincingly.

**2. The raw and extracted columns compare a model tuned on the output format against models that were not.** FinGEITje scores identically before and after extraction on four of five Dutch tasks (SA 0.790, HC 0.920, RE 0.569, QA 0.324), while extraction lifts FinMA from 0.276 to 0.632 accuracy on sentiment and from 0.091 to 0.792 F1 on NER (CLS) (Table 5, F1 row; the accuracy row for the same cell moves from 0.236 to 0.776). That is the signature of a format effect rather than of financial understanding, and it means a share of the gap measures how well each system happens to emit the expected label string. A format-agnostic protocol applied to every system, for instance constrained decoding or log-likelihood scoring over the label set, would isolate the part of the margin that comes from the financial data.

**3. Two of the five Dutch wins sit inside sampling noise.** Evaluation uses 500 samples per task, so the 95% interval on a difference between two accuracies in this range is roughly six points. Sentiment analysis (0.790 against 0.752 for gpt-3.5-turbo-0125, both taken from the extracted column; on the raw column, the gap is 4.8 points) and question answering (0.324 against 0.294 for FinMA after extraction) fall inside that band; headline classification, NER, and relation extraction are far outside it and are safe. No seeds, repetitions, or intervals appear anywhere in the paper, and since Section 4.3 claims victory on all tasks, it is the weakest cell that decides the wording.

One more point belongs here. The benchmark is translated by the same model, with the same prompt, as the training data (Appendix A), so both sides carry the same translationese. Section 3.1.3 names that risk and then leaves it unmeasured, although it works directly in favour of a model trained on that distribution and against GPT-3.5 and the general Dutch models. A few hundred natively Dutch or human-translated items would convert the caveat into a measurement.

### *Reproducibility*

Better than most work of this kind. The base model, the adapter method (QLoRA over all linear layers), batch size 8, learning rate 2.0e-04, 10% warmup, one epoch, two A100 40GB GPUs, FlashAttention, and the API snapshot gpt-3.5-turbo-0125 with temperature 0 and max_tokens 1024 are all stated, and both prompts are printed in Appendices A and B. That is more than the field usually offers.

What is still missing to re-run it: random seeds, both for training and for the draw of the 500 evaluation samples; LoRA rank, alpha, and dropout; the quantisation used; maximum sequence length; gradient

accumulation and therefore the effective batch size and precision; and the decoding settings at inference, that is, greedy or sampled, temperature, top-p, maximum new tokens, stop sequences. That last gap matters more than it looks, because the raw column is, in effect, a measurement of format compliance, which decoding settings alone can move. It is also unclear whether the financial-expert persona from Section 3.1.2 was used at evaluation or only during data construction.

Finally, gpt-3.5-turbo-0125 is a snapshot on its way out. Once it is withdrawn, neither the translation nor the answer extraction can be reproduced as described. Archiving the raw generations together with the extractor inputs and outputs would keep these numbers checkable after the endpoint disappears.

### *What to fix*

1. Section 3.3 and Table 4: state explicitly how training and evaluation data were separated and at which stage of the pipeline; add a near-duplicate check over the final split and report the overlap rate.
2. Section 4.1.2 and Appendix B: validate the extractor. Hand-label a random sample (200 items is enough), report agreement and per-model error rates. Extraction moves FinMA sentiment by 35.6 points and FinGEITje by zero, so extractor error is not neutral with respect to the ranking.
3. Section 4.1.1: give the evaluation protocol per model, that is decoding parameters, prompt template, and whether the English baselines received Dutch or English instructions. Few-shot baselines would make the comparison fairer, since in-context examples are the standard remedy for format failures.
4. Section 4.2: report confidence intervals, or at minimum the per-cell sample size, and soften the claim in Section 4.3 to name the tasks where the margin is larger than noise.
5. Section 3.1.3: quantify translation quality on a sample, with adequacy and fluency ratings by two annotators and an agreement statistic, plus the label-flip rate on sentiment items, where a translation can move the gold label.
6. Section 3.1.4: define the term Signifiers; it is used as if it were established. State also the granularity of exact deduplication, that is instruction, input, response, or the whole triple.
7. Add the ablation the paper needs: the same base tuned on an equally sized translated non-financial instruction set. Without it, the financial data and the output format cannot be told apart as causes.
8. Table 1: FinTral is attributed to [17], while the text attributes it to [14]; reference [17] is FinMA-ES. FinGPT likewise appears as [4] in the table and [13] in Section 2.2.

9. Table 2: NWGI and Finance Alpaca carry no citation and no licence. Given the copyright statement in Section 7, per-dataset licences for the translated derivatives, tweets in particular, should be listed.
10. Table 5 against Section 4.1.1: the column is labelled geitje-7b-chat while the text names GEITje-7B-ultra-sft. Align the naming, or say which checkpoint was actually run.
11. Footnote 3: the URL merges two namespaces, huggingface.co/Rijgersberg/BramVanroy/GEITje-7B-ultra-sft, and does not resolve.
12. Small text fixes: Section 6explores in Section 1 is missing a space, and the Conclusion says comprising over more than 140,000 samples, which says it twice; Table 3 gives 147,788, so quote that figure.

### *Recommendation*

Major revisions. Points 1 to 3 decide whether the headline comparison means what it is said to mean, and all three can be answered with work the authors have already half done: a statement about the split plus a duplicate check, one uniform scoring protocol, and intervals on the numbers. Nothing here questions the contribution itself. Even on the narrower claim that survives, namely that for Dutch financial tasks in this instruction format a 7B model tuned on machine-translated financial data beats general Dutch models and zero-shot GPT-3.5, the release remains a useful piece of infrastructure, and the translation pipeline is the part other languages will reuse.

Evgenii Arsentev, PhD

## Declarations

**Potential competing interests:** No potential competing interests to declare.