

REVIEWER'S REPORT

Manuscript No.: IJAR-59512

Title: How Reliable Are Layer Comparisons from Small Probing Datasets?

Recommendation:

Accept as it is

Accept after minor revision.....

Accept after major revision yes.....

Do not accept (*Reasons below*)

Rating	Excel.	Good	Fair	Poor
Originality			Y	
Techn. Quality			Y	
Clarity			Y	
Significance			y	

Reviewer's ID: JPR-Dr.Shaweta Sachdeva

Detailed Reviewer's Report

1. The authors should justify the use of 512 sentences as the reference training budget and clarify whether the selected layer remains stable as the training sample increases.
2. The small audit sample, particularly the three LinES audit documents, raises concerns about the reliability of the uncertainty estimates. The authors should discuss this limitation more explicitly.
3. The authors should provide confidence intervals for the warning metrics, where feasible, to better demonstrate the reliability of the reported AUROC, precision, and recall values.
4. The observed reduction in selection errors with increasing Qwen3 model size is interesting; however, the effects of parameter count, representation width, layer depth, and checkpoint training should be distinguished more carefully.
5. The absence of above-tolerance errors in 30 decisions for the larger Qwen3 models should be interpreted cautiously and should not be taken as evidence that selection errors are negligible.
6. The study considers only three English treebanks and two token-level classification tasks. The authors should discuss the generalizability of the findings to other languages, linguistic properties, and probing tasks.
7. The authors should clarify the rationale for selecting five depths per model and discuss whether evaluating a subset of layers affects the conclusions regarding layer selection.
8. The manuscript should better explain how lexical memorization may influence probing accuracy and how the word-type control experiments support the interpretation of linguistic information encoded in model representations.
9. The authors should provide a stronger justification for the regularization grid and discuss whether the selected range is appropriate for models with different representation dimensions.
10. The performance of geometric selection methods varies across model checkpoints. The authors should clarify whether the findings reflect limitations of the specific implementations evaluated or broader limitations of geometry-based selection approaches.
11. The authors should distinguish more clearly between changes in the selected layer and practically meaningful accuracy losses, emphasizing the role of the predefined one-percentage-point tolerance.

International Journal of Advanced Research

Publisher's Name: Jana Publication and Research LLP

www.journalijar.com

REVIEWER'S REPORT

12. The statistical analysis would benefit from paired uncertainty estimates when comparing selection methods, particularly because the methods share datasets, training subsets, and evaluation settings.
13. The authors should clarify the rationale for using sentence-averaged accuracy as the primary evaluation metric and discuss how it affects comparisons across sentences of different lengths.
14. The manuscript reports changes to the experimental protocol, including corrections to the development-data partition and additional experiments introduced after earlier results were examined. These changes should be documented clearly to facilitate assessment of potential analytical bias.
15. The authors should provide complete implementation details, including checkpoint identifiers, configuration files, and analysis scripts, to improve the reproducibility of the experiments.
16. The discussion should more clearly distinguish the ability of warning scores to rank potentially unreliable selections from their ability to identify errors using thresholds transferred across datasets.
17. The authors should clarify the definitions and relationships among winner changes, reference regret, oracle regret, and excess error to improve the manuscript's readability.
18. The conclusion should provide more specific practical guidance on when small-sample layer selection can be considered reliable and when additional training data or independent validation may be necessary.