

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 ...✓.....

Do not accept (*Reasons below*)

Rating	Excel.	Good	Fair	Poor
Originality		✓		
Techn. Quality		✓		
Clarity			✓	
Significance		✓		

Reviewer's ID: JPR- **098**

Detailed Reviewer's Report

Decision: Accept after Major Revision

Major Revisions

1. Strengthen the justification for the selected training budgets.

The study compares probing datasets containing 32, 128, and 512 sentences. Although these budgets provide a useful experimental progression, the manuscript should justify why these sample sizes were selected and explain how representative they are of practical low-resource probing scenarios. The authors should also discuss whether the observed selection errors persist at intermediate or larger training budgets.

2. Address the limited diversity of linguistic datasets.

The experiments use three English treebanks: EWT, GUM, and LinES. These datasets provide useful variation, but they do not establish that the findings generalize to other languages, language families, annotation schemes, or linguistic tasks. The authors should expand the evaluation if feasible or explicitly limit their conclusions to the examined English treebanks and tasks.

REVIEWER'S REPORT

3. Investigate the effect of the very small audit samples.

The development-data audit partitions contain 65 documents for EWT, 11 for GUM, and only 3 for LinES. Such differences can substantially affect the precision of uncertainty estimates and the usefulness of the proposed warning methods. The authors should discuss the implications of these sample sizes, provide sensitivity analyses where possible, and avoid interpreting the failure of warning scores as definitive evidence that such methods are generally ineffective.

4. Clarify the interpretation of the one-percentage-point error threshold.

The manuscript defines selection regret greater than one percentage point as an error exceeding the chosen tolerance. However, this threshold is a practical criterion rather than a statistical significance test. The authors should explain its practical motivation, distinguish it clearly from statistical significance, and discuss whether the principal conclusions change under alternative thresholds. The reported 0.5- and 2-percentage-point sensitivity analyses should be incorporated more prominently into the main discussion.

5. Improve statistical uncertainty reporting.

The manuscript reports mean regret, confidence intervals for some comparisons, and variation across experimental settings. However, the uncertainty associated with the main comparisons could be explained more systematically. The authors should clarify how the reported intervals were calculated, what observations they treat as independent, and whether clustering by model, treebank, task, or random seed affects their interpretation. Where appropriate, paired comparisons or cluster-aware uncertainty estimates should be considered.

6. Examine the potential influence of model architecture and context availability.

The study compares bidirectional encoders, causal decoders, and an encoder-decoder model. These models differ not only in architecture but also in the contextual information available to each token, pretraining procedures, model size, and checkpoint characteristics. The authors acknowledge some of these

REVIEWER'S REPORT

limitations, but should discuss them more prominently and avoid attributing performance differences solely to architecture.

7. Strengthen the interpretation of the Qwen3 model-size experiment.

The manuscript reports substantial selection errors for Qwen3-0.6B compared with the other examined checkpoints. However, model size, representation width, depth, and checkpoint training vary together. Consequently, the results do not establish that model size alone causes the observed differences. The authors should clarify this distinction and discuss the additional controlled experiments needed to isolate individual factors.

8. Clarify the fairness and interpretation of the layer-selection comparisons.

The study compares ridge tuning accuracy, nearest-centroid classification, geometric scores, and a final-layer baseline. These methods differ in their use of labels, hyperparameter selection, and access to validation information. The authors should explain the information available to each selector and ensure that comparisons reflect clearly defined and practically relevant selection settings. The manuscript should also distinguish retrospective evaluation using the larger training budget from information available when making the original low-data selection.

9. Improve the discussion of lexical memorization and linguistic information.

The word-type control experiments and lexical lookup baseline are valuable additions. However, the ability to predict arbitrary labels associated with word types does not, by itself, quantify the amount of memorization or syntax encoded by a language model. The authors should explain the limits of these controls, discuss how word identity and contextual representations interact, and avoid interpreting control-task performance as a direct measure of linguistic knowledge.

10. Strengthen reproducibility and the practical contribution of the study.

The manuscript provides considerable methodological detail, including training budgets, seeds, feature standardization, regularization candidates, and model checkpoints. Nevertheless, the reproducibility of the complete experiment would be improved by providing accessible code, exact data-

REVIEWER'S REPORT

sampling procedures, the full configuration of each selection rule, and scripts for reproducing the tables and figures. The authors should also clarify how practitioners should choose a layer when the larger-data reference results are unavailable in real-world applications.

Minor revisions

1. Refine the abstract. The abstract should state the main quantitative findings more concisely, including the frequency of layer changes, mean selection regret, and the concentration of larger errors in particular checkpoints.
2. Improve the statement of novelty. The introduction should distinguish the paper's contribution more explicitly from earlier research on probing sample efficiency, layer selection, and representation-based model-selection methods.
3. Standardize mathematical notation. Ensure consistent notation for layer indices, training budgets, validation accuracy, test accuracy, regret, and learning gains throughout the equations and accompanying explanations.
4. Clarify the terminology for selection errors. The terms “reference regret,” “excess error,” “oracle regret,” “winner change,” and “ranking reversal” should be defined consistently and distinguished clearly in the methods and results sections.
5. Improve table readability. Table headings should identify units and clarify whether values represent percentages, percentage-point differences, or proportions. Explain the meaning of the reported “Match” column and ensure that the table notes are self-contained.
6. Improve figure captions and cross-referencing. Each figure caption should identify the training budget, evaluation metric, selection method, and experimental subset represented. Ensure that all figures are cited and discussed in the main text.
7. Clarify the averaging procedure. The manuscript states that model, treebank, task, and seed combinations receive equal weight. Explain this aggregation procedure and distinguish macro-averaged results from token-level or sentence-level accuracy.

International Journal of Advanced Research

Publisher's Name: Jana Publication and Research LLP

www.journalijar.com

REVIEWER'S REPORT

8. Correct minor grammatical and typographical issues. Carefully proofread the manuscript for awkward sentence constructions, inconsistent hyphenation, spacing around mathematical symbols, and formatting errors introduced during equation or table conversion.

9. Improve the organization of the theoretical section. The propositions concerning label dependence, coordinate scaling, and independent audit bounds are relevant. However, the transition from mathematical results to empirical experiments could be more direct, with a brief statement of the practical implication of each proposition.

10. Standardize references and supplementary materials. Verify the completeness and accuracy of all references, including model and dataset documentation. Ensure that the appendices contain the exact layer depths, complete selection results, sensitivity analyses, and precision comparisons referenced in the main text.