Explanation of Resubmitting Revisions

The authors would like to thank the editor and reviewers for their constructive comments and suggestions that have helped improve the quality of the paper "**LM-LEXICON: Improving Definition Modeling via Harmonizing Semantic Experts**", which has been submitted to the ARR of July 2025. The resubmitted manuscript has undergone a thorough revision according to the AC and reviewers' comments. Please see the responses below. For reviewers' convenience, we have highlighted significant changes in the revised manuscript in RoyalBlue.

Area Chair

Meta Review — This paper proposes a novel method for definition modeling called LM-Lexicon that uses sparse MoE to combine semantic expert training. Experimental evaluations over baselines on multiple datasets show the benefits of the proposed model. While reviewers agree about the relevance of the studied task and promising results, they express several concerns that need to be addressed.

There ia a summary of suggested revisions:

(1) Clearly state the novelty of the contribution beyond BTX.

(2) Clarity of the methodology: better define variables and equations. Provide more detail about clustering (e.g., why only cluster 3D-EX) and add details about the training setup.

(3) Human evaluation is not clear, and the justification for why it is only conducted on a proprietary model is missing. Provide also more details on human evaluation.

(4) Better analysis of the use of LLMs and the report training setup.

Reply: We thank the Area Chair for these suggestions. We have addressed them in the revised manuscript as follows:

- Novelty
- Clustering details
- Human evaluation
- Training setup

Please refer to the following replys for more details.

Reviewer A

Reviewer bYwM

Reviewer Comment A.1 — Some general weaknesses:

1. In equation (1) the probability of $\hat{d}$ is optimised, however this variable has not been introduced, according to the appendix the model is optimised for the gold definition d. The expectation is defined over every (p, c, t) tuple, it seems that the gold definition d and not the static prompt function p should be specified. Furthermore, the indicator function is defined over a single token, but multiplied with the whole sequence, which appears to be inconsistent with the intended definition. It might be clearer to restructure this equation.

2. Could the authors clarify why human trials were performed only over the proprietary model results, when Rerank T5 would have been a stronger comparison based on the results in Table 2.

3. The description of Table 3 states that it compares results in WordNet over various settings, however the base BLEU and ROUGE scores don't match the WordNet results in Table 2. They do however match the 3D-EX results. Is this a typo?

Reply: We thank the reviewer for the comments. We have carefully revised the original manuscript in response to each of the reviewer's suggestions.

1. We have restructured Equation 1 into

$$\mathcal{L}_{\mathsf{NLL}} = -\,\mathbb{E}_{(c,t,d)\sim\mathcal{D}}\left[\log \mathcal{P}(\hat{d} \mid p(c,t))\right]. \tag{1}$$

   where $\hat{d}$ denotes the predicted definition.

   - We use the standard negative log-likelihood loss as the learning objective, in which the expectation is defined over every (c, t, d) tuple.
   - The loss is only computed on tokens in the definition part not the prompt. We have clarified this setup in the manuscript (see in L205-L209).

2. Our primary objective in the human trials was to evaluate the effectives and practicability of various methods. While we acknowledge that Rerank-T5 shows strong performance in Table 2, the leading performance in this domain is currently dominated by such proprietary systems, making them the most relevant benchmarks for assessing practical impact. That said, we agree that a comparison with Rerank-T5 in human evaluation might also be insightful. To address this, we have added the rationale in the corresponding section of the manuscript (Section 4.2, Paragraph "Human Evaluation", L368-L372).

3. We have corrected typos, specifically "Wordnet" to "3D-EX" in the relevant positions. We apologize if this notation caused confusion, and we have clarified this in the text.

Reviewer Comment A.2 — Model training: The paper does not clearly state what data the final model is trained over. It lists five training datasets in section 4.1 'Datasets', giving the description: "We perform clustering only for 3D-EX and use the other four datasets as the natural clusters for the training and merging of semantic experts." In later sections it is explained that four clusters proved optimal for 3D-EX, this suggests that the model was trained over eight clusters in total (four "natural" dataset clusters and four clustered 3D-EX splits). However the final model in Table 2 is reported to have 4x8B parameters, suggesting only four experts/clusters. Could you specify the training data setup more clearly?

If the model was trained on all five datasets, were any precautions taken to prevent test set leakage? As shown in Figure 5, all other datasets are at least partially represented in the large 3D-EX training set.

Reply: We thank the reviewer for raising these important questions regarding our training data setup and potential data leakage.

Our training process involves two distinct phases:

Phase 1 (Individual Expert Training): We perform instruction tuning separately on each data cluster. This includes the four clusters derived from the 3D-EX dataset via clustering and the four natural, independent datasets (WordNet, Oxford, Wikipedia, Urban). This results in eight individual expert models. The results of the models trained on four independent datasets are reported in LM-LEXICON-DENSE (8B) in Table 2. The four experts from the natural datasets remain separate all the time.

Phase 2 (Merging and Final Finetuning): The model architecture after merging consists of **only** four experts representing the merged 3D-EX knowledge. The reported model size (4x8B) refers to the MoE model with 4 experts obtained from 3D-EX. The final model is then finedtuned over the full 3D-EX training set.

We confirm that these details, particularly the two-phase training process (individual tuning followed by merging and router fine-tuning), are described in Section 4.1, Paragraph "Training and Evaluation Details." To enhance clarity, we have revised the text in this section to more explicitly state the the number of experts in the final MoE model and to reiterated the strict separation between training and evaluation across all dataset to prevent any leakage.

Reviewer Comment A.3 — Baselines: There are uncertainties regarding the specifics of how baseline values were calculated. The paper states: "To compare with these baselines broadly, we replicate the setups used by prior work and reuse their reported results whenever possible." However in the final results of Table 2 every listed paper is marked with a clover, indicating it was reproduced using supervised training. Could the paper provide more details regarding which results were reused? Analogous to model training, it is not quite clear which data is used during baseline replication and evaluation, could this be specified? If baseline models are retrained, it might be helpful to list the training cost in the carbon footprint section.

Unrelated to this, there are some small details missing, such as which model was used for the LlamaDictionary baseline (Llama2 or Llama3), and what prompts were used when querying the proprietary LLMs.

Reply: We thank the reviewer for the comments.

- Replication details: We have correct the inappropriate expression "reporting their original results whenever possible", instead, we reproduce all of the baselines and report the experimental results conducted by us. The reproducing code and implementaion will be released upon completion of the review process. For each baseline replication, we use the same training dataset and test dataset for them, to keep consistent with LM-Lexicon for fair comparison.

- Model used in LlamaDictionary baseline: We have already specify the base model as **"Llama-3-8B-instruct"** in the Section 4.1 in the text. For better clarification, we hightlight the base model in bold fonts. We use the same prompt template for querying all baselines including proprietary models as well as the LM-Lexicon.

Reviewer Comment A.4 — Human evaluation: I would be curious to know more details about the human evaluation, such as for instance if the evaluators had access to model names and the original dictionary definitions. An estimate of annotator agreement might also be interesting.

Reply: Human evaluators have no access to model names to avoid possible bias, whereas the reference definitions remained accessible to evaluators. These details have been added to the manuscript in Section 4.2, Paragraph "Human Evaluation" for clarity.

Reviewer B

Reviewer 1imX

Reviewer Comment B.1 — The proposed model is basically Branch-Train-MiX (BTX). Although its application to definition modeling is valuable, the novelty of the proposed model is limited. To obtain semantic experts, the paper introduces k-means clustering, which is straightforward. Another contribution is the introduction of domain-level routing, which is inspired by previous work (L441-442). However, it is hard to understand its method because its details are not described.

Reply: We thank the reviewer for her/his feedback and for acknowledging the value of applying our approach to definition modeling. While our work is indeed inspired by the general Branch-Train-Mix (BTX) paradigm, our key contribution lies in a distinct "Cluster-then-Train" methodology, which differs significantly from the typical BTX workflow. Instead of branching before training on potentially mixed data, our approach first clusters the training data (specifically 3D-EX) based on semantic similarity to identify distinct domains. We then train specialized experts on these pre-defined semantic clusters.

The core novelty resides in the **domain-level routing** strategy we introduce for the final merged MoE model. The method is as follows:

i. Semantic domain representation: Each semantic domain (cluster) is represented by its centroid, derived from the initial k-means clustering of the training data embeddings.

ii. Routing mechanism: For an input x, its embedding is computed. The router calculates the pairwise cosine similarity between this input embedding and the centroid embedding of each predefined semantic domain.

iii. Expert selection: The input x is routed to the expert corresponding to the semantic domain (centroid) with the highest cosine similarity. This effectively means the input is routed to the "nearest" semantic domain centroid.

We agree that the initial description could be clearer, and we have revised the explanation of this routing mechanism in Section 3.2.

Reviewer Comment B.2 — Some variables are not defined in the equations in Section 3.

Reply: We have revised the equations in Section 3 to make them clearer. Please refer to the revised manuscript for more details.

Reviewer Comment B.3 — Comments Suggestions And Typos:

- L187: ”context” should be defined precisely.
- L247: domains.Similarly should be domains. Similarly
- Table 2: ”Slang” should be ”Urban” as defined in Section 4.1.

Reply: We thank the reviewer for carefully reading and poiniting these typos out. Corrections have been made in the revised manuscript.

Reviewer C

Reviewer iWL9

Reviewer Comment C.1 — We doubt the research significance of definition modeling, as the real world does not require so many new definitions.

Reply: Language is envolving in nature, and the call for the dictionary or lexicon in new form is neccesary for language evolution. Hence, we argue that definition modeling is still an important task to investigate.

Reviewer Comment C.2 — Comparing T5-base baselines with llama3-based LM-Lexicon is not fair enough.

Reply: We have already listed more baselines in the original paper. Except for the T5-like seq2seq architecture, we also include a transformer decoder-based causal language model (LlamaDictionary) which is trained on Llama-3-8B-instruct.

Reviewer Comment C.3 — The experimental setting of clustering only for 3D-EX and use the other four datasets as the natural clusters for the training and merging of semantic experts is unreasonable. You should merge the five datasets and then perform clustering.

Reply: We have clarified the clustering procedure in Section 4.1. It is clear enough to see why we conduct this different setups for 3D-EX and other four datasets.

Reviewer Comment C.4 — It is uncertain whether this method is applicable to LLMs other than Llama 3.

Reply: We believe that training based on Llama-3 is sufficient to demonstrate the empirical effectiveness and efficiency of our work.

Reviewer Comment C.5 — The training cost is high, as there are four llama3-8b that need to be trained.

Reply: We think that performing full parameter training is a better way to incentivizing the semantic modeling capability of LM-Lexicon.