

Dear Action Editors and Reviewers,
 We thank the Action Editors and Reviewers for their thoughtful, constructive feedback. Below we respond point-by-point. We present the reviewers concern in *red italics*, followed by our response in black; *blue excerpts* are verbatim insertions added to the revised manuscript.

General Clarification. We have gone through every comment in details and we feel that majority of the concerns have stemmed from the general misunderstanding of the position of the original manuscript. It seems to come across, quite understandably, that DegreeD is the key contribution of the paper and acts as an objective function for PerAugy. However, we have tried to clarify both in this response and also in the revised manuscript (we have decoupled discussion on DegreeD and other diversity metrics to Sec 7) that **PerAugy remains the key contribution and not DegreeD**. PerAugy has been designed to induce diversity in original datasets via Double Shuffling (DS) and Stochastic Markovian Perturbation (SMP) operations. Since there is no other operation than DS and SMP involved while keeping every other setting the same (including dataset size), *the performance gain of models trained on PerAugy-augmented datasets cannot possibly be attributed to anything other than dataset diversity*. **This is already established from the results in Sec 6: Tables 2-3 and App I.1: Fig 6 & 8.** Hence, DegreeD (like other simpler diversity metrics like TP and RTC) is **rather a post-hoc diagnostic metric** to quantify the diversity that has been induced by PerAugy and other baseline augmentation methods. As mentioned later, **any analysis of DegreeD (or any other such metric) is more of a matter of which diversity metric is better suited to quantify such induced diversity in a reliable manner.**

Reviewer qSYo

Concern-1: Stronger Support for DegreeD-Accuracy Correlation

Response: We thank the reviewer for pointing out this valid concern. As per the suggestion, we strengthened the evidence on two fronts in Sec 7.3: (i) stability w.r.t PerAugy as a strong outlier (Tables 6) and (ii) stability w.r.t strong user-encoder outlier (Table 7) across all datasets (both original and augmented). We find that **DegreeD (along with one of the other diversity metrics, TP) tracks model accuracy robustly even when excluding PerAugy-augmented datasets**. Also, we observe that DegreeD and TP are stable w.r.t. strong positive user-encoder outlier, and hence, the aggregate mean correlation reported in Table 6 is reliable. We

also give a detailed mathematical analysis on the stability of DegreeD’s correlation results upon substitution of σ as RMSD with other alternative σ (App. F, p. 37, L36-49; p. 38, L17).

Concern-2: Circularity of PerAugy and DegreeD.

Response: We completely acknowledge the misunderstanding that has been raised across the reviewers. We admit that the paper’s narrative structure led to that, and hence, **we have restructured the revised manuscript by keeping DegreeD and other diversity metrics in a separate discussion (Sec 7) on quantification of the induced diversity of PerAugy (and other augmentation methods).** We would like to clarify that PerAugy is designed to increase *trajectory diversity* in UIGs, not to optimize DegreeD directly (i.e., as an optimization objective). The DS operation, which is key to diversity induction, does not explicitly model the three types of penalties (unfaithfulness, disproportionate divergence, and lack of divergence) baked into DegreeD. In other words, PerAugy is agnostic to the ratio-based alignment of document vs. summary shifts that DegreeD measures. To emphasize, DegreeD is a model/performance-metric-agnostic **post-hoc diagnostic dataset diversity metric** to *quantify (and not prove) the role of dataset diversity in performance gains* (Sec. 7.1). This makes **PerAugy the key contribution of the paper, and not DegreeD** (which is merely a diversity metric like TP and RTC). This addresses circularity by decoupling PerAugy’s mechanism from the evaluator.

Concern-3: DegreeD’s Sensitivity to Choice of Embedding & Divergence Metric.

Response: We understand the concern and acknowledge that this has not been clearly treated in the original manuscript. We clarify that DegreeD computation depends on *relative* ratio rather than absolute scales (App. E, p. 33-34, L36-42: "*any embedding model ... and any valid distance metric σ (e.g., Manhattan, cosine, Euclidean) merely provide a representation space ... the ratio-based structure of DegreeD cancels out biases due to embedding geometry or metric scaling. Hence, DegreeD remains a model- and metric-agnostic measure ...*"). We also clarify the default setting (App. E, p. 33, L47-50): "*we represent d/s-nodes with S-BERT embeddings ... and use Manhattan Distance as σ* ". At the same time, we provide a theoretical analysis that establishes the stability of DegreeD under divergence metric substitution (App. F, p. 37, L36-49; p. 38, L17): "*Theorem 1 (Correlation stability of DegreeD) ... Scaling invariance ... preserved exactly.*". We, however, would like to emphasize that DegreeD (and any other arbitrary dataset diversity metric) is only important to the extent of strengthening the *interpretability* of what the effect of DS and SMP operations is w.r.t diversity. In fact, even

the simpler TP metric shows the same trend, further confirming that **our findings do not hinge on one specific metric design – it is more of a question of which is a better diagnostic metric rather than a conclusion on the performance of PerAugy itself.**

Concern-4: Clarification of DegreeD Notations.

Response: We simplified and standardized notation and consolidated all symbols in Table 1 (p. 15). Clarifications include explicit superscripts / subscripts for δ across document vs. summary channels and introduction of the regulator α (p. 15, L19): " *α Regulator that controls the influence of Penalized DePS.*". We also tightened the surrounding prose in Sec. 7.1.

Concern-5a: Example-based Clarification of DePS; Concern-5b: Clarification of min-max ratio.

Response: We added worked examples and intuition in Sec. 7.1. The min/max design yields bounded, scale-invariant proportionality, aligning with the ratio-based robustness emphasized in App. F (bi-Lipschitz stability) and Sec. 7.1.

Concern-6: Elaboration of Related Work.

Response: We expanded Sec. 8 (Related Work).

Concern-7: Confusion between *genSumm* and *summGen*.

Response: We clarified the UIG actions explicitly in Sec. 3 (p. 4, L48 & L52). Inserted text: *summarize (also called genSumm) explicitly captures the interest to read a summarized version of the d-node and "the follow-up edge of summarize denoting summarized version of d_{t_q-1} , acting on s_{t_q} (also termed as summGen)*". (Sec. 3, p. 4, L48-52; Fig. 3)

Concern-8: Regarding Sufficiency of Test Data.

Response: We describe the construction and scale of the PENS test setup used for next-click and summarization evaluation. The test pool is substantial: "*The final test set contains 103 trajectories with $\approx 20K$ candidate pool ... of d-nodes with 10K target d-nodes.*" (Sec. 5.1.1, p. 9, L52-54) This PENS test settings is largely adopted in literature "PENS: A Dataset and Generic Framework for Personalized News Headline Generation, Ao et. al.", "General then Personal: Decoupling and Pre-training for Personalized Headline Generation, Song et.al." (Table 8).

Concern-9: Usage of Simpler TP/RTC Metrics.

Response: We do a head-to-head comparison of TP, RTC, and DegreeD, both theoretically and empirically. Table 4 shows that TP/RTC correlate with diversity but fail to penalize mis-aligned summary shifts. Caption (p. 16, L17): "*While TP and RTC relate to diversity, they fall short in capturing preference ...*". DegreeD explicitly measures proportional alignment

between document and summary drift, which best tracks personalization accuracy (Sec. 7.3).

Concern-10: Regarding Reference Style

Response: We have corrected it throughout (there was a newer TMLR package compatibility issue).

Reviewer hXxr

Concern-1: Regarding Sensitivity to Hyperparameters & Overfitting Risk.

Response: We added a comprehensive ablation suite (App. I.1-I.2). Sec. I.1 details RQ-1 ablations over DS gaplength g_ℓ , train history length, and SMP parameters k , λ , p_{SMP} (p. 39, L51-55). Sec. I.2 summarizes DegreeD ablations with concrete settings (p. 41, L27-29) – "*optimal setup being $k = 10$, $\lambda = 0.3$, and $p_{SMP} = 1$ (with a score of 0.278)*". These results show robustness windows rather than knife-edge tuning. Also, models trained in PerAugy-augmented OpenAI-Reddit dataset (augmented under the same configurations as PENS augmentation) have been tested in the PENS test dataset, showing promising results. This also indicates that underfitting/overfitting cannot be attributed to the hyper-parameters. Detailed results are given in Sec 6.1 (Cross-domain study) and corresponding Figure 8 (Appendix I; pg: 42).

Concern-2: Regarding Coverage beyond PENS dataset.

Response: We agree that this is a very valid concern. However, we would humbly like to point out that we have done cross-domain analysis leveraging OpenAI (Reddit) data. This dataset is a non-news set of Reddit threads covering 29 different domains, as compared to the PENS dataset (Sec 3.2, Sec 5.1.1: Training Data). OpenAI-Reddit is also structurally different than Reddit. We find diversity gains and accuracy trends persist (Sec 6: Cross-domain-study & App. I: Fig. 8).

Concern-3: Regarding Comparative Study with Simpler Interpretable Metrics.

Response: We thank the reviewer for pointing out this very important discussion that was omitted in our last submission. In the revised version, we do a head-to-head comparison of TP, RTC, and DegreeD, both theoretically and empirically. Table 4 shows that TP/RTC correlates with DegreeD. Also, we find that the PerAugy-augmented dataset diversity across all three metrics (DegreeD, TP, RTC) shows an increase when compared to the corresponding original dataset. **This helps us to consistently quantify the induced diversity by PerAugy.** We also discuss at length how TP

and RTC can fail to penalize misaligned summary shifts but at the same time are more interpretable. Caption (p. 16, L17): "*While TP and RTC relate to diversity, they fall short in capturing preference ...*". DegreeD explicitly measures proportional alignment between document and summary drift, which best tracks personalization accuracy (Sec. 7.3). We would finally like to stress that *our findings on the performance of PerAugy do not hinge on one specific metric design* – it is more of a question of **which is a better post-hoc diagnostic metric of the induced diversity (not just by PerAugy but by any other augmentation method), rather than a conclusion on the performance of PerAugy itself.**

Concern-4a: Stronger Support for DegreeD-Accuracy Correlation; Concern-4b: Degreed-PerAugy Tight Coupling.

Response: We admit that the paper’s narrative structure led to the misunderstanding of DegreeD being tightly coupled to PerAugy. Hence, **we have restructured the revised manuscript by keeping DegreeD and other diversity metrics in a separate discussion (Sec 7) on quantification of the induced diversity of PerAugy (and other augmentation methods).** We would like to clarify that PerAugy is designed to increase *trajectory diversity* in UIGs, not to optimize DegreeD directly (i.e., as an optimization objective). The DS operation, which is key to diversity induction, does not explicitly model the three types of penalties (unfaithfulness, disproportionate divergence, and lack of divergence) baked into DegreeD. In other words, PerAugy is agnostic to the ratio-based alignment of document vs. summary shifts that DegreeD measures. To emphasize, DegreeD is a model/performance-metric-agnostic **post-hoc diagnostic dataset diversity metric** to *quantify (and not prove) the role of dataset diversity in performance gains* (Sec. 7.1). This makes **PerAugy the key contribution of the paper, and not DegreeD** (which is merely a diversity metric like TP and RTC).

As for the suggestion regarding strengthening the correlation results, we have done that on two fronts in Sec 7.3: (i) stability w.r.t PerAugy as a strong outlier (Tables 6) and (ii) stability w.r.t strong user-encoder outlier (Table 7) across all datasets (both original and augmented). We find that **DegreeD (along with one of the other diversity metrics, TP) tracks model accuracy robustly even when excluding PerAugy-augmented datasets.** Also, we observe that Degreed and TP are stable w.r.t strong positive user-encoder outlier, and hence, the aggregate mean correlation reported in Table 6 is reliable. We also give a detailed mathematical analysis on the stability of DegreeD’s correlation results upon substitution of σ as RMSD with other alternative σ (App. F, p. 37, L36-49; p. 38, L17).

Concern-5: Regarding Reference Formatting.

Response: We have corrected it throughout (there was a newer TMLR package compatibility issue).

Concern-6: Regarding Qualitative Examples of Augmented Trajectories.

Response: We added concrete DS/SMP examples and case studies-Fig. 3(b,c) and Sec. 4.2 include explicit trajectory snippets (p. 7, L20-21).

Concern-7: Limitations without ground-truth histories.

Response: We have discussed this issue in Sec. 9 (p. 20, L18): "*model-generated summaries can serve as effective proxies, especially during cold-start scenarios ...*". We also show how OpenAI-Reddit styled datasets can be appropriated to facilitate bootstrapping (Sec. 3) – "*Additionally, it also addresses cold-start problem as $\mathcal{T}_{base}^{syn-OAI}$ itself is synthetically designed as a random sequence*".

Reviewer silk

Concern-1: Robustness of Correlation Relation & Generalizability Across Alternative Augmentation Methods.

Response: We agree that (i) correlations alone do not prove causation and (ii) extremely noisy or misaligned augmentation could in principle harm learning. Two sets of results in our paper address these concerns directly:

(A) Baseline augmentations also induce diversity and (often) improve accuracy (not just PerAugy). We evaluate three baseline algorithmic augmentors (PENS-SH, S3-Aug, SDAInter) and three LLM-based augmentors (LLaMA-2, Mistral, DeepSeek). As summarized in Table 2, **several baselines improve encoder accuracy relative to the original PENS training set (e.g., SDAInter yields consistent gains across NAML, EBNR, NRMS; PENS-SH substantially lifts TrRMIo under fine-tuning).** This shows that increased (and well-aligned) diversity is beneficial even when PerAugy is not used. In contrast, others are neutral or harmful, depending on how their induced diversity aligns with trajectory semantics. This, conversely, shows that misaligned or low-quality diversity can hurt, as pointed out by the reviewer – i.e., too diverse (or noisy) data degrading learnability. To make this explicit in the paper, we already enumerate the baseline methods and the way each induces diversity in UIGs (inter-trajectory merges in PENS-SH, intra-trajectory perturbations in S3-Aug, and cross-user subsequence swaps in SDAInter) before measuring their effects on encoders. All the augmentation methods increase dataset diversity w.r.t the three diversity metrics (TP, RTC, and DegreeD) – see Table 4.

(B) Correlation between diversity and accuracy holds across all datasets, and remains strong even when PerAugy is excluded.

The meta-evaluation in Sec 7 computes correlations between diversity metrics (TP, DegreeD) and encoder accuracy across the entire set of original and augmented datasets (including baselines). Table 6 shows that DegreeD and TP both correlate positively with accuracy; critically, Observation-3 reports that these correlations remain strong even after excluding PerAugy, so the trend is not because of PerAugy dominating the pool. This directly counters the concern that only PerAugy goes up and drags correlation - the relationship between (well-measured) diversity and accuracy persists for other augmentation families too. We also run stability checks (Section 7.3.3): removing PerAugy (outlier-method analysis) leaves the correlation band essentially unchanged (reported mean deltas and low variance), and analyzing model-specific correlations shows low inter-model variance, with DegreeD generally more stable than TP. Together, these indicate that the diversityaccuracy trend is robust and not driven by a single method or a single encoder.

To conclude, baseline augmentations do induce diversity and often improve accuracy, and the positive diversityaccuracy association holds broadly, not only with PerAugy. Where some baselines underperform, our results support a ***quality-of-diversity*** explanation (semantic misalignment / noisy diffusion) rather than a blanket "*too much diversity causes underfitting.*" PerAugy’s DS+SMP is designed precisely to regulate diffusion (DS) and enforce faithfulness/temporal coherence (SMP), which explains its stronger and more consistent gains. (We already highlight that DS-only is weaker than DS+SMP across encoders.)

Concern-2: Regarding the Necessity of SMP Operations.

Response: We thank the reviewer for pointing this out. We saw that we omitted the result earlier and have included both DS-only vs. DS+SMP comparisons on both encoder and summarizer tasks (Tables 2-3) in the revised manuscript. As noted in Table 3 caption (p. 13, L11-12): "*SMP consistently improves over DS, indicating its necessity for maximizing gains.*". Downstream ablations are given in App. I.1; Fig 6, while results are discussed in Sec 6.1.

Concern-3: Example-based Clarification of DS Operation.

Response: We inserted a concrete stitched example in Sec. 4.2 with explicit operations (p. 7). For instance (p. 7, L20-21): "*Alices first three interactions involve reading “Meditation tips” ... and “Yoga retreat guides” ...*".

Concern-4: DegreeD’s Sensitivity to Choice of Embedding &

Divergence Metric.

Response: We understand the concern and acknowledge that this has not been clearly treated in the original manuscript. We clarify that DegreeD computation depends on *relative* ratio rather than absolute scales (App. E, p. 33-34, L36-42: "*any embedding model ... and any valid distance metric σ (e.g., Manhattan, cosine, Euclidean) merely provide a representation space ... the ratio-based structure of DegreeD cancels out biases due to embedding geometry or metric scaling. Hence, DegreeD remains a model- and metric-agnostic measure ...*"). We also clarify the default setting (App. E, p. 33, L47-50): "*we represent d/s-nodes with S-BERT embeddings ... and use Manhattan Distance as σ* ". At the same time, we provide a theoretical analysis that establishes the stability of DegreeD under divergence metric substitution (App. F, p. 37, L36-49; p. 38, L17): "*Theorem 1 (Correlation stability of DegreeD) ... Scaling invariance ... preserved exactly.*". We, however, would like to emphasize that DegreeD (and any other arbitrary dataset diversity metric) is only important to the extent of strengthening the *interpretability* of what the effect of DS and SMP operations is w.r.t diversity. In fact, even the simpler TP metric shows the same trend, further confirming that **our findings do not hinge on one specific metric design – it is more of a question of which is a better diagnostic metric rather than a conclusion on the performance of PerAugy itself.**

Concern-5a: Notation Clarification of δ notation; Concern-5b: Missing Notation Brief in Symbol Table.

Response: We fixed notation in Sec. 7.1 and updated Fig 2b Table (p. 15), including the " *α Regulator that controls the influence of Penalized DePS.*" Symbols for δ across timesteps are now explicit.

We hope the revised manuscript resolves the concerns and improve clarity, rigor, and reproducibility.