The revision of this submission mainly focus on the following three parts:

1 Empirical Validation of Privacy-Utility Tradeoff

Addressing the meta-reviewer's primary concern about "lack of corresponding evaluation for how query utility is actually affected in practice by masking PII", we have added a comprehensive section empirically evaluating this tradeoff using 200 unique user descriptions from PII-Bench. We systematically compared three masking strategies: No Mask (baseline), All PII Mask (maximum privacy), and Query-unrelated PII Mask (our proposed strategy). Results demonstrate that our Query-unrelated PII Mask achieves an optimal balance between privacy protection ($P = 0.83$) and utility preservation ($U = 0.89$). We also added an appendix formalizing our evaluation framework with: (1) privacy protection metrics using entity occurrence counts, (2) utility preservation measurements via embedding-based semantic similarity (BGE-M3) and LLM-as-Judge evaluation (Claude-3.7-Sonnet), and (3) a combined metric integrating privacy and utility scores. This analysis quantifies necessary privacy-utility tradeoffs and evaluates variations across model capabilities.

2 Evaluation of Smaller Models

Following reviewer recommendations, we evaluated smaller deployment-ready models (Qwen2.5-0.5B, Qwen2.5-1.5B, and Qwen2.5-3B) across all PII-Bench datasets using various inference strategies (Naive, Self-CoT, Auto-CoT, Self-Consistency, PS-CoT). Results reveal a clear scaling trend, with the 3B model showing reasonable performance while smaller models exhibit significant degradation, providing insights for on-device privacy protection deployment.

3 Real-world Dataset Validation

To address concerns regarding data realism, we constructed and evaluated a new real-world dataset called PII-Real containing biographical profiles of 20 renowned AI scholars from the AMiner AI2000 ranking. We manually annotated PII entities according to our guidelines, generated corresponding queries following our established methodology, and created a validation set of 100 instances.

We are currently conducting comprehensive experiments to evaluate model performance on PII-Real across various architectures and scales.