Accelerating Personalization Signal Learning via Synthetic Data

Daraksha Parveen✉, Doug Kang, Anwitha Paruchuri, Deep Kayal, and Pavan Mallapragada

Alexa AI, Amazon
{dxparvee,dougkang,paruchua,dkayal,pavanmns}@amazon.com

Abstract. Personalized experiences in multimodal assistants rely on accurate user understanding, yet large-scale training for personalization remains limited by privacy constraints and data sparsity. We introduce a framework for generating *Comprehensive Synthetic Personas* (CSPs) and personalized synthetic training data through taxonomy-guided knowledge enrichment, in-context learning, and Chain-of-Thought (CoT) knowledge distillation for personal knowledge inference. This dataset is used to fine-tune a large language model (LLM) that learns to infer user interests and attributes from interaction histories. Evaluation shows that models trained on synthetic data outperform those trained on real de-identified data in precision and recall when evaluated on the same real de-identified test set, using LLM-as-a-Judge (LLMaaJ) and human annotation. This approach enables scalable and privacy-safe personalization learning, supporting downstream applications in AI assistants.

Keywords: Personal Knowledge Inference · Synthetic Data Generation · Comprehensive Synthetic Personas · LLM-as-a-Judge Evaluation

1 Introduction

Learning personalization signals is critical for adaptive user experiences in large-scale industry assistants. However, large-scale supervised training is constrained by privacy restrictions, data sparsity, and limited topical diversity in real world production interactions [7,2]. To address these challenges, we introduce a synthetic, privacy-preserving pipeline for personal knowledge inference, enabling end-to-end model training without reliance on real user data. Our framework constructs comprehensive personas that capture behavioral, demographic, and interest attributes, and leverages large language models (LLMs) with taxonomy-guided knowledge enrichment and in-context learning to synthesize realistic, persona-grounded dialogues paired with structured personal knowledge items (PKIs). The training dataset is used to fine-tune an LLM for personalization learning through Chain-of-Thought (CoT) knowledge distillation [8,16]. Evaluation is performed using an LLMaaJ assessing precision, QA-based recall, and human assessment for PKI accuracy. Models trained exclusively on synthetic data achieve gains of +5% in precision and +8% in recall for PKIs compared to baselines trained on real world de-identified customer data. Customer annotation results further confirm over 85% acceptance, demonstrating the effectiveness of synthetic personas for scalable, privacy-safe personalization and their applicability to downstream AI assistants interest modeling.

2 Related Work

Open-domain conversational agents frequently elicit personal information implicitly during free-form interactions, motivating research on inferring user attributes for personalization and user modeling. Early work on persona-conditioned dialogue generation, notably Persona-Chat [27], showed that conditioning responses on persona descriptions improves coherence and engagement, while also demonstrating that speaker attributes can be inferred from dialogue histories. Synthetic-Persona-Chat [11] uses LLMs to generate large-scale, faithful dialogues, addressing limitations of earlier persona datasets in diversity, consistency, and scalability. PersonaChatGen [15] introduced taxonomy-driven persona representations for automated persona based dialogue generation. SODA [13] shows that grounding synthetic dialogue generation in structured knowledge graphs improves semantic diversity and coherence, while PLACES [3] shows that few-shot expert exemplars can effectively guide LLM-based dialogue generation. Other works formalize personal attribute inference as a standalone task. Wu et al. [23] introduce an early distant-supervision framework with structured triple prediction, while CHARM [20] enables zero-shot inference of unseen attribute values via contextual cues and retrieval. More recent methods emphasize structured representations and generalization. GenRe [21] separates explicit extraction from implicit inference using a generator–reranker architecture, while PAED [28] improves data quality and evaluates generalization to unseen attributes. SynthPAI [26] introduces a large-scale synthetic benchmark for personal attribute inference demonstrating alignment between synthetic and real-data evaluation outcomes.

3 Methodology

3.1 Synthetic Data Generation Pipeline

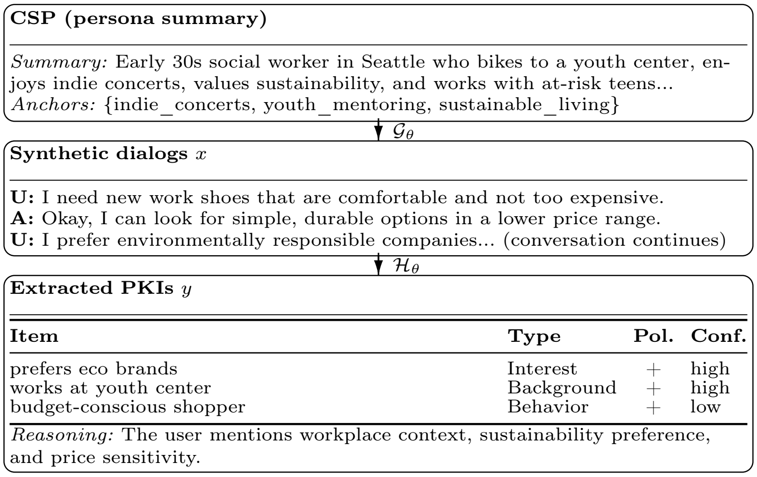
Comprehensive Synthetic Personas We construct *Comprehensive Synthetic Personas* (CSPs) to serve as structured representations of user behavior, interests, and background attributes. The persona profile schema consists of a unified hierarchical ontology of seven top-level tiers spanning domains such as interests, backgrounds, and behaviors. Each tier $t \in \mathcal{T}_{\text{tier}}$ defines a set of categories $\mathcal{C}_t$, and each category $c \in \mathcal{C}_t$ contains a collection of entity types $\mathcal{E}_{t,c}$. The full schema is given by $\Omega = \bigcup_{t \in \mathcal{T}_{\text{tier}}} \{(t, c, e) \mid c \in \mathcal{C}_t, e \in \mathcal{E}_{t,c}\}$, where each triple (t, c, e) denotes a distinct attribute key instantiated with a short natural-language value (e.g., “Likes Taylor Swift” under the *Interest* tier, the *Arts & Entertainment* category, and *Favorite Musician* entity type). In total, the schema defines $|\Omega| = 130$ attributes per persona, alongside a unique identifier and a natural-language persona summary describing the overall profile.

Persona Generation and Knowledge Enrichment CSPs are generated through a non-parametric prompting process that combines LLM reasoning with taxonomy-guided knowledge enrichment. The enrichment step samples from the Interest Taxonomy $\mathcal{T} = (V, E)$, a directed tree of over 1800 interest nodes organized hierarchically across 25+ categories (e.g., one path is Academic Interests & Careers → Humanities → Philosophy). From $\mathcal{T}$, we retrieve a set of semantic anchor paths $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, where each path $p_i = (v_i^{\mathrm{root}} \to v_i^1 \to \cdots \to v_i^{k_i})$ connects the root to a sampled leaf node. Terminal nodes $v_i^{k_i}$ are selected using a hybrid strategy that blends random and depth-stratified sampling to balance topical diversity with semantic specificity. Given the enriched paths $\mathcal{P}$, the schema Ω, and the instruction set $\mathcal{I}$, a foundation model $\mathcal{M}_\theta$ generates a filled profile $U \leftarrow \mathcal{M}_\theta(\mathcal{I}, \mathcal{P}, \mathit{\Omega})$. Full taxonomy paths ground attributes within semantic subspaces of $\mathcal{T}$ and act as interpretable interest dimensions that guide the model to generate diverse but contextually consistent persona attributes.

Personalized Training Dataset Generation We generate synthetic dialogs that simulate realistic user interactions grounded in each synthetic persona U. A role-playing model $\mathcal{G}_\theta$ produces multi-turn conversations conditioned on the persona, sampled semantic paths $\mathcal{P}$ as conversation seeds, and few-shot exemplars of real AI assistants deidentified utterances for in-context learning. The model receives structured instructions to vary linguistic realizations across affective dimensions such as *intensity* (high, medium, low) and *polarity* (positive, negative). These dimensions are defined by distributional targets derived from production interaction statistics, ensuring generated utterances reflect realistic behavioral variation. An LLM-based extractor $\mathcal{H}_\theta$ then identifies and annotates PKIs from each dialog with metadata including type, polarity, intensity, and reasoning traces (see Fig. 1) following a structured decomposition-based thinking framework [22]. The resulting dataset comprises conversations, extracted PKIs, reasoning traces, persona summaries, and is used for chain-of-thought instruction fine-tuning [14] of downstream personalization models.

3.2 Synthetic Data Validation Methodology

To evaluate the quality of the synthetic dataset, we conduct a comparative analysis [17] by training LLMs on (1) synthetic data generated from our CSP framework, and (2) real de-identified production data. All models are evaluated on a real production de-identified test set to quantify performance and verify that synthetic data does not degrade generalization to real-world interactions [24,19]. Performance is measured using LLM-as-a-Judge (LLMaaJ)–based precision and recall metrics and human evaluation ratings [25]. We also analyze failure modes and edge cases where synthetic data may introduce artifacts or spurious correlations absent in real customer conversations [1]. For fine-tuning, we utilize LLMs with parameters $\leq$ 10B, balancing performance with production deployment constraints such as inference latency, and GPU capacity ([12],[9]). We conduct experiments using two training techniques on both datasets to generate PKIs: Low-Rank Adaptation (LoRA) [10] and Supervised instruction fine-tuning.



CSP (persona summary)

Summary: Early 30s social worker in Seattle who bikes to a youth center, enjoys indie concerts, values sustainability, and works with at-risk teens...
Anchors: {indie_concerts, youth_mentoring, sustainable_living}

Synthetic dialogs x

U: I need new work shoes that are comfortable and not too expensive.
A: Okay, I can look for simple, durable options in a lower price range.
U: I prefer environmentally responsible companies... (conversation continues)

Extracted PKIs y

Item	**Type**	**Pol.**	**Conf.**
prefers eco brands	Interest	+	high
works at youth center	Background	+	high
budget-conscious shopper	Behavior	+	low

Reasoning: The user mentions workplace context, sustainability preference, and price sensitivity.

Fig. 1. High-level overview of personalized training dataset generation

Precision For precision, we measure the proportion of correctly extracted PKIs among all extracted PKIs, evaluated with the LLMaaJ framework for *relevance* and *sensibility*. Relevance assesses whether a PKI is explicitly stated or can be reliably inferred from the customer's interaction history with sufficient evidence (for e.g., "I love pop music" → *pop music*; repeated "Play Taylor Swift" requests → *Taylor Swift*). Sensibility assesses whether the PKI reflects a meaningful user interest rather than a routine requests or an illogical association. A PKI is considered precise when both criteria are satisfied. The reliability of this metric is validated through human annotation, yielding a Cohen's κ of 0.62.

Recall Measuring recall is challenging since the ground truth of customer interests is rarely observable. We adopt a QA-based recall metric [6], used in summarization evaluation where reference completeness is similarly difficult to guarantee [5,4]. A questionnaire aligned with customer use cases is manually designed, and an LLM generates answers to these questions using customers' interactions (*gold* answers) and inferred PKIs (*candidate* answers). Candidate responses are compared against gold references: "I don't know" answers are marked incorrect (TP=0, FN=1) when a gold answer exists, while valid pairs are scored via Sentence-BERT cosine similarity [18], with scores above a threshold counted as true positives and those below as false negatives. Outputs are manually reviewed to ensure evaluation reliability, as correlation with human judgment is non-trivial.

4 Results

We evaluate model performance using both automated metrics and customer validation. Models fine-tuned on the persona-based synthetic dataset outper-

Table 1. Performance comparison across training configurations on a real world de-identified test set. Metrics are percentage-point differences relative to a >20B-parameter baseline. All experimental methods use the same <10B base LLM with different adaptation strategies

Training Data	Adaptation Method	*Δ*Precision	*Δ*Recall
None (baseline, >20B)	Zero-shot prompting	0.0%	0.0%
None (<10B)	Zero-shot prompting	-3.4%	+4.8%
Real world de-identified	LoRA	+2.0%	+1.0%
Real world de-identified	Full SFT	+3.0%	+4.9%
Synthetic	LoRA	+3.5%	+4.9%
Synthetic	Full SFT	**+8.0%**	**+13.0%**

Table 2. Models trained on synthetic data outperform models trained on real world de-identified data across fine-tuning strategies on a real de-identified test set (mean over 3 runs)

Fine-tuning	*Δ*Prec.(%)	*Δ*Recall (%)
LoRA	+1.5	+3.9
Full SFT	**+5.0**	**+8.0**

Table 3. Synthetic data statistics compared to real world de-identified data

Metric	Synthetic
Avg. turns per dialog	+2.2
Vocabulary diversity	+4.6K
PKIs per dialog	+0.9
Interest category coverage	+31%

form those trained on real de-identified data, achieving +5% precision and +8% QA-recall gains as shown in Table 1 and Table 2. For customer validation, a representative sample of 115 users reviewed their learned PKIs generated by the synthetic-persona-based model through an internal annotation tool. Users assessed the accuracy of each PKI, yielding an overall precision above 85%, demonstrating strong user acceptance and system effectiveness.

5 Conclusion

We present a framework for generating CSPs and personalized training data via taxonomy guided knowledge enrichment, in context learning, and role playing generation. The synthetic dataset enables privacy preserving fine tuning of personalization models without reliance on real user data. Evaluation using LLMaaJ and human validation shows substantial gains in precision and recall for PKIs, with over 85% customer validated precision, demonstrating the effectiveness of synthetic data for scalable industry scale personalization. Empirically, we find that taxonomy guided persona grounding over a 1,800 node interest ontology preserves semantic diversity, mitigates overfitting to synthetic artifacts, and improves generalization across both head and long tail domains that are underrepresented in real logs. In addition, CSPs provide fully specified personal knowledge supervision including polarity, confidence, and reasoning traces, yielding a denser learning signal that supports effective chain of thought distillation and improves performance on real world de-identified data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bayer, M., Kaufhold, M.A., Reuter, C.: A survey on data augmentation for text classification. ACM Computing Surveys **55**(7), 1–39 (2022)
2. Bukharin, A., Li, S., Wang, Z., Yang, J., Yin, B., Li, X., Zhang, C., Zhao, T., Jiang, H.: Data diversity matters for robust instruction tuning. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3411–3425. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.195, https://aclanthology.org/2024.findings-emnlp.195/
3. Chen, M., Papangelis, A., Tao, C., Kim, S., Rosenbaum, A., Liu, Y., Yu, Z., Hakkani-Tur, D.: PLACES: Prompting language models for social conversation synthesis. In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 844–868. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.findings-eacl.63, https://aclanthology.org/2023.findings-eacl.63/
4. Deutsch, D., Roth, D., Durrett, G.: Towards question-answering as an automatic metric for evaluating the content quality of summaries. Transactions of the Association for Computational Linguistics **9**, 346–361 (2021)
5. Durmus, E., He, H., Diab, M.: Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 5055–5070 (2020)
6. Fabbri, A.R., Wu, C.S., Liu, W., Xiong, C.: QAFactEval: Improved qa-based factual consistency evaluation for summarization. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 2587–2601. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.naacl-main.187, https://aclanthology.org/2022.naacl-main.187/
7. Fu, X., Chen, N., Gao, P., Li, Y.: Privacy-preserving personalized recommender systems. SSRN Working Paper (2022), https://ssrn.com/abstract=4202576
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS Deep Learning Workshop (2015)
9. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 30016–30030 (2022)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
11. Jandaghi, P., Sheng, X., Bai, X., Pujara, J., Sidahmed, H.: Faithful persona-based conversational dataset generation with large language models. arXiv preprint arXiv:2312.10007 (2023)
12. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)

13. Kim, H., Hessel, J., Jiang, L., West, P., Lu, X., Yu, Y., Zhou, P., Le Bras, R., Alikhani, M., Kim, G., Sap, M., Choi, Y.: SODA: Million-scale dialogue distillation with social commonsense contextualization. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12930–12949. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.emnlp-main.799, https://aclanthology.org/2023.emnlp-main.799/
14. Kim, S., Joo, S.J., Kim, D., Jang, J., Ye, S., Shin, J., Seo, M.: The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. arXiv preprint arXiv:2305.14045 (2023), available at https://arxiv.org/abs/2305.14045
15. Lee, Y.J., Lim, C.G., Choi, Y., Im, J.H., Choi, H.J.: Personachatgen: Generating personalized dialogues using gpt-3. In: Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge. vol. 1, pp. 29–48 (2022)
16. Magister, L., Fried, D., Belinkov, Y.: Teaching small models to reason: Distilling chain-of-thought (2023), https://arxiv.org/abs/2305.06350
17. Nikolenko, S.I.: Synthetic Data for Deep Learning, Springer Optimization and Its Applications, vol. 174. Springer (2021)
18. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3982–3992 (2019)
19. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 1–48 (2019)
20. Tigunova, A., Yates, A., Mirza, P., Weikum, G.: Charm: Inferring personal attributes from conversations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). vol. 1, pp. 5391–5404 (2020)
21. Wang, Z., Zhou, X., Koncel-Kedziorski, R., Marin, A., Xia, F.: Extracting and inferring personal attributes from dialogue. In: Proceedings of the 4th Workshop on NLP for Conversational AI. vol. 1, pp. 58–69 (2022)
22. Wen, P., Ji, J., Chan, C.M., Dai, J., Hong, D., Yang, Y., Han, S., Guo, Y.: Thinkpatterns-21k: A systematic study on the impact of thinking patterns in llms. arXiv preprint arXiv:2503.12918 (2025), https://arxiv.org/abs/2503.12918
23. Wu, C.S., Madotto, A., Lin, Z., Xu, P., Fung, P.: Getting to know you: User attribute extraction from dialogues. In: Proceedings of the 12th Language Resources and Evaluation Conference (LREC). vol. 1, pp. 581–589 (2020)
24. Xu, C., et al.: How does synthetic data generation impact machine learning performance? a comprehensive study. arXiv preprint arXiv:2301.09286 (2023)
25. Yao, Y., et al.: Evaluating the text generation capabilities of large-scale language models. arXiv preprint arXiv:2207.07411 (2022)
26. Yukhymenko, H., Staab, R., Vero, M., Vechev, M.: A synthetic dataset for personal attribute inference. In: Advances in Neural Information Processing Systems (2024), https://arxiv.org/abs/2406.07217
27. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). vol. 1, pp. 2204–2213 (2018)
28. Zhu, L., Li, W., Mao, R., Pandelea, V., Cambria, E.: Paed: Zero-shot persona attribute extraction in dialogues. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). vol. 1, pp. 9771–9787 (2023)