PersonaBase: Computational Resources for Advancing the Research and Practice of Data-Driven Personas

Joni Salminen
University of Vaasa
Vaasa, Finland
jonisalm@uwasa.fi

Danial Amin
University of Vaasa
Vaasa, Finland
danialam@uwasa.fi

Ilkka Kaate
University of Turku
Turku, Finland
iokaat@utu.fi

Bernard J. Jansen
Peking University
Beijing, China
jjansen@acm.org

Abstract

Researchers struggle to replicate studies and measure progress in data-driven persona (DDP) research, while practitioners struggle to understand, develop, and validate DDPs using shared methods and metrics. To address these challenges, we adopt ideas from the natural language processing (NLP) community to present PersonaBase, a repository of computational DDP resources that gathers tasks, datasets, metrics, and baselines. The first version of PersonaBase contains 11 notebooks demonstrating algorithmic persona generation and evaluation approaches, 27 persona generation prompts for large language models (LLMs), 19 systems, and 12 datasets to experiment with. We describe these DDP resources and evaluate target users' first-impression utility with two formative studies, one with 4 persona researchers and another with 25 industry practitioners. The evaluators generally found the resources useful and relevant, with more experienced DDP users identifying datasets and LLM prompts as particularly relevant. Both researchers and practitioners exhibited needs for (1) improved documentation and onboarding, (2) reproducible pipelines, (3) template-based resources, and (4) improved accessibility for non-technically oriented users. We propose that future work on DDP resources addresses these needs by community building and learning from other computational sciences, while moving closer to full benchmarking solutions.

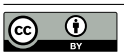
CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Data-driven personas, benchmarking, computational resources

ACM Reference Format:
Joni Salminen, Danial Amin, Ilkka Kaate, and Bernard J. Jansen. 2026. PersonaBase: Computational Resources for Advancing the Research and Practice of Data-Driven Personas. In *Proceedings of the 14th Nordic Conference on Human-Computer Interaction (NordiCHI '26), October 03–07, 2026, Vaasa, Finland*. ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3829807.3829819



This work is licensed under a Creative Commons Attribution 4.0 International License.
NordiCHI '26, Vaasa, Finland
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2373-5/26/10
https://doi.org/10.1145/3829807.3829819

1 Introduction

> "*I would like to call into question the idea that [benchmarking] will support reproducibility. At least in machine learning and AI generally, automated evaluation, benchmarks, and fixed metrics became targets to optimise rather than measures of quality, which has resulted in chasing SOTA numbers rather than actually producing useful insights. Persona creation has been deeply qualitative, and I worry that the fate of 'benchmarkification' at the expense of qualitative insights would await this area if metrics are introduced.*" –Anonymous reviewer

> "*[...there is a] need to develop a rigorous science of persona generation and outline the methodological innovations, organizational and institutional support, and empirical foundations required to enhance the reliability and scalability of [...] persona simulations*" –Li et al. [48] (p. 1).

The former quote above from an anonymous reviewer on an earlier version of this work illustrates the concerns the human-computer interaction (HCI) community has about developing computational techniques for data-driven persona (DDP) development in seeing those techniques risking the emergence of qualitative insights. In contrast, the latter quote from a computationally oriented research paper articulates the opposite risk: that the lack of methodological rigor can harm persona research. Therefore, there exist two parallel narratives: one drawing toward more quantitative persona generation, another toward maintaining qualitative "ancestry" of persona development. The tension between these narratives foregrounds the present work, which is focused on how computational resource development could benefit DDP research.

In user-centered design (UCD), representing real user groups through fictional *personas* [17] is a well-established methodology to stimulate empathy and guide decision making between developers, designers, and other stakeholders who need information about users [25, 63]. Persona is a tool for UCD by providing an empathetic view of users based on data instead of an abstract idea of "user" based on vague assumptions [17]. Although personas were initially created

using small-scale qualitative data [26], scholars adopted large-scale datasets to develop DDPs, which are based on quantitative data analysis [36, 53, 56]. The progress in natural language processing (NLP) and generative artificial intelligence (GenAI) has broadened the scope of DDP development to leverage unstructured data, such as user study transcripts, interviews, and/or open ended surveys that can be processed using GenAI techniques to develop persona descriptions [3, 19, 71, 82].

The increasing reliance of persona generation on computational techniques anchors the DDP research domain strongly in computational research, at least method-wise. Despite this anchoring, DDP development currently lacks systematically developed computational resources and practices to establish quality standards and methodological rigor similar to other computational fields [91, 92]. Systematic reviews have found that less than 10% of persona studies share resources such as code, datasets, or algorithms [75]. The DDP research community has repeatedly called for the improvement of DDP research rigor; e.g., Chapman et al. in 2006 [12] and 2008 [11], Mijac et al. [56] in 2018, Salminen et al. [78] in 2022, and Li et al. in 2025 [48], among others. To enhance rigor, one crucial direction is the development of *computational resources*, which are digital artifacts like source code, datasets, programming notebooks, algorithms, and models. Without shared computational resources, it is difficult to systematically evaluate and improve DDPs within the research community.

While this lack of computational resources for DDP development hinders replicating results, it also makes it more difficult to identify where (and by how much) scientific progress DDP research field has made. By scientific progress, we mean the ability of a research community to produce knowledge and tools that *improve* specific, commonly agreed upon metrics by building on top of previous work [8, 70]. Better user understanding and methodological rigor through empirical testing, replication, and building on previous work are important for the HCI research community. Research shows that shared resources increase progress in computational sciences [7, 91]. For example, in NLP, benchmarks such as GLUE [92] and SuperGLUE [91] facilitate scientific progress by allowing comparisons between competing algorithms and models. The DDP research domain would similarly benefit from standard resources to guide the development and evaluation of DDPs, particularly because persona research is strongly affected by the increasing use of data, algorithms, and computational models, such as large language models (LLMs) [18, 38, 71, 80, 82, 97] that introduce novel risks, such as hallucination and other algorithmic biases [18, 38, 71, 80, 82, 97]. While introducing novel challenges, technology development also introduces opportunities for the development of DDP resources; for example, shared prompt libraries [3]. Against this backdrop, the current study addresses the following research question: *How to develop computational resources to advance research on DDPs?*

Addressing this research question provides the HCI research community with a situational awareness of the current state of computational resources in DDP research. Regarding the scope of our work, the reader should note that this article is not written for NLP experts or machine learning (ML) specialists to whom this content may appear "self-evident" and "trivial" (or even uninteresting!). The paper *is* written for HCI researchers and practitioners to provide useful definitions of the basic concepts of benchmarking and to explain how they could be valuable in DDP development and evaluation. For this reason, throughout the manuscript, we provide several contextual definitions in the footnotes to help the reader better understand the benchmarking terminology.

Our contribution focuses on introducing and providing a formative first-impression evaluation of the `PersonaBase (PB)`, a computational resource repository on DDPs (see Figure 1). `PB` aims to advance DDP research and practice by curating shared datasets, GenAI prompts, and example computational notebooks for DDP development and evaluation. `PB` targets two user groups. For *novice researchers* seeking to get started with DDPs, `PB` offers practical example notebooks that demonstrate ML algorithms in DDP development, along with datasets that help setup experiments. Overall, `PB` helps users learn the basics of DDP development by applying some of the resources in their own use case. For *expert researchers*, `PB` offers ideas for a shared foundation for benchmarking new methods against existing approaches using standardized datasets and metrics, for which we provide example code. Ideally, `PB` enables experts to build on previous work while providing diverse datasets to test new techniques and contribute to the development of common standards in DDP research.

As a key limitation (more of which we discuss in Section 5.1), we do not claim that `PB` addresses benchmarking for DDPs in the same sense as GLUE or SuperGLUE did for NLP. Our contribution is closer to a curated resource base and a proposal for how the field might gradually move toward benchmarking.

In the remainder of this work, we will first offer the reader a perspective on the evolution of DDPs in HCI research. Then, we discuss benchmarking principles from the NLP domain, and how these are crucial "stepping stones" toward computational resource integration also in DDP research. We then aim to bring clarity and structure into the fragmented, somewhat chaotic landscape of computational resources in DDPs. In this process, we articulate how `PB` provides value in addressing the gaps we identify and conduct formative evaluation studies with persona researchers and practitioners. Finally, we discuss the implications and the future research directions that we believe would benefit the DDP research community.

2 Conceptual Underpinnings

2.1 Data-Driven Personas as Computing Research

While persona research in HCI has traditionally been characterized by qualitative efforts [60–63], persona methods have shifted to a quantitative direction over the years, alongside the evolution of data science [56, 78]. However, this qualitative tradition has meant that persona studies often deal with small sample sizes (i.e., 5-20 participants), involve interpretative methodologies, and focus on idiosyncratic findings rather than generalizable patterns [4]. While NLP researchers working on persona-based dialogue systems routinely publish replication studies using standardized datasets [50, 95], HCI researchers rarely re-use datasets, develop their own algorithms, or apply the same evaluation metrics.

Initially, DDPs were introduced as alternatives to generate, validate, and deploy personas with greater efficiency and scalability

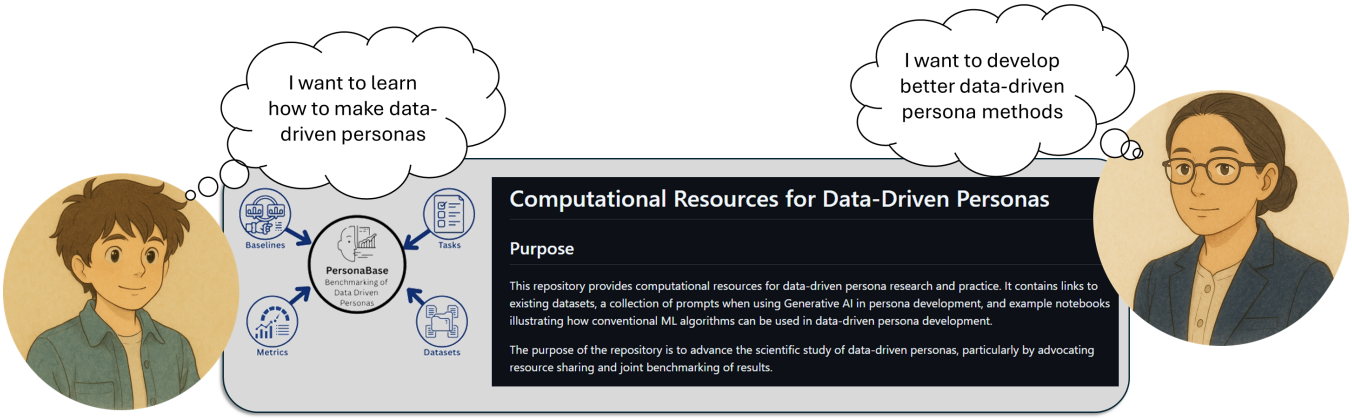


Figure 1: PersonaBase represents a partial solution to reproducibility and benchmarking problems in DDP research, targeting the needs of both novice and more experienced persona developers. The repository entails computational resources for learning and further developing DDP methods. The repository is available on GitHub: https://github.com/joolsa/PersonaBase. Disclaimer: the two illustrative personas appearing in the figure have been generated using GPT-5.

compared to manual methods, which often rely on qualitative work inputs that were considered expensive, slow, and non-tractable compared to computational techniques [4, 37, 53]. Research on personas has evolved to involve more and more computational techniques. Reviewing this evolution, Salminen et al. [75] identified three eras in DDPs: (1) emergence, (2) diversification, and (3) sophistication. Of these, the *Emergence* era (2005-2008) introduced quantitative methodologies for persona research, whereas the *Diversification* era (2009-2014) expanded the methodologies to include NLP and behavioral data while developing greater self-awareness of the DDP field. The *Sophistication* era (2015-2020) involved increasingly complex applications, including interactive persona systems (e.g., Automatic Persona Generation (APG) [37]), persona analytics, and real-time persona creation from web and social media data.

In addition to these eras, we could identify the fourth era, starting in 2022 with the introduction of LLMs, possibly dubbed the *Interaction* era, in which LLMs increasingly contribute to persona development, evaluation, and modes of providing the personas to their users [1, 2, 38, 99]. The application of LLMs has also enabled persona-based simulations for decision-making scenarios in user research, economic modeling, social science, etc. [31, 48]. Overall, the evolution of DDPs is increasingly aligning DDPs with computational research practices and traditions, which implies that meta-scientific methods from other computational domains, such as NLP, can be of value in DDP development (see Figure 2 for an illustration). Following this logic, we next review some of the key ideas behind computational resource development in NLP.

2.2 Computational Resources in Natural Language Processing

Because benchmarking practices are highly developed in NLP, it is logical to draw inspiration from practices in NLP when discussing the development of computational resources for DDPs. We therefore provide an overview of how NLP approaches computational resources, as we believe this to be useful information to the reader. This information also supports the development of our main argument, which is that computational resources are useful for measuring progress in DDP research. To this end, the NLP community has demonstrated the *value of establishing shared resources*, most notably common (1) tasks (e.g., GLUE[1] [92], SuperGLUE[2]) [91], (2) datasets (e.g., SQuAD[3] [72]), (3) evaluation metrics (e.g., BLEU[4] [68], ROUGE[5] [49]), and (4) baselines (i.e., "proven methods" that generally work well for a given problem type). The idea with baselines is, roughly speaking, that if a method does not outperform the simplest known solution, then it is not worth being implemented. In addition to baselines, a key consideration is state-of-the-art (SOTA) performance. To be justified as a valuable contribution, a proposed new solution should surpass the SOTA performance. Through these commonly known baselines and SOTA performance values, the NLP community sets a *shared norm* that requires researchers to show quantitative evidence of better performance.

In NLP, computational resources create a shared vision and *scientific understanding* to compare methods, measure progress, promote reproducibility, and guide newcomers to ways to contribute. This vision is *collective* and *legitimate* among the NLP research community; researchers know about it, accept it, and act according to it. Evidence of how these shared resources drive NLP forward can be seen, for example, in the fact that the transformer model introduced by Google researchers [20] was eventually adopted and

[1] GLUE (General Language Understanding Evaluation) is a benchmark used in NLP to evaluate how well language models understand and process human language across multiple tasks.
[2] A more challenging benchmark designed to address limitations of the original GLUE benchmark as models began to approach or exceed human performance on the original tasks.
[3] Stanford Question Answering Dataset (SQuAD) is a reading comprehension resource featuring over 100,000 questions created by crowdworkers based on Wikipedia articles. It was developed to test natural language understanding and question answering capabilities in AI systems.
[4] BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of machine-translated text by comparing it to human reference translations.
[5] ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate automatic summarization and machine translation by comparing an automatically produced summary or translation against reference human-written versions.

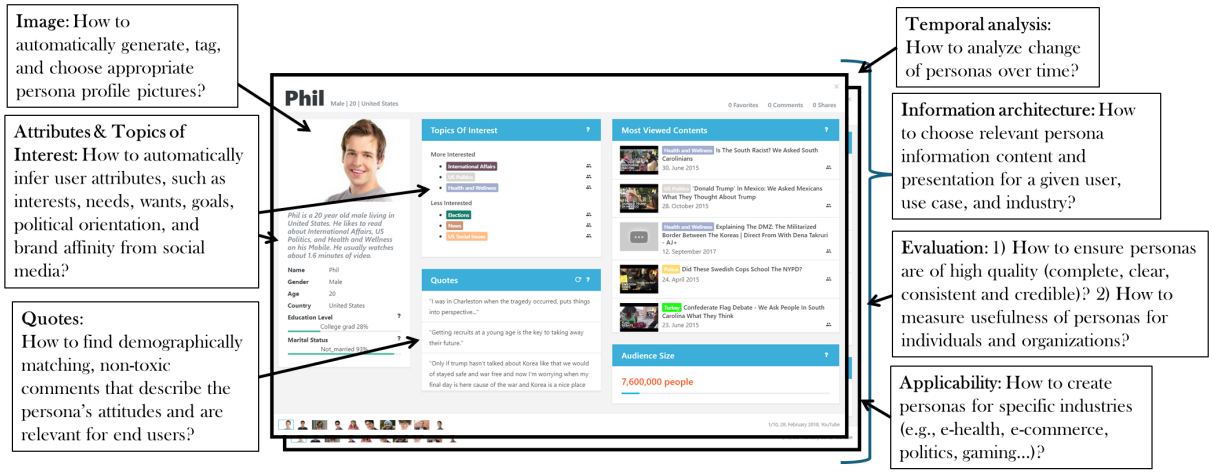


Figure 2: Illustration of how challenges in persona profile creation can be translated into computational tasks (the illustration uses the Automatic Persona Generation system [40]). Various elements of a typical persona profile can be formulated as computational tasks. If shared metrics are provided for these tasks, researchers can propose their unique solutions to addressing each task. However, DDP research has not thus far defined tasks and metrics with adequate granularity.

further developed by OpenAI's researchers, after which GenAI became a mainstream technology that is currently transforming most domains of intellectual activity. Consequently, DDP research would benefit from similar resources to make measurable progress in developing and validating DDPs.

Benchmarking in NLP tends to be based on a few core components that establish a framework to objectively evaluate scientific contributions (see Table 1). In general, benchmark datasets define the *tasks* the models are tested on, and *metrics* define how model performance is measured on those tasks [90]. In NLP benchmarking, a task is a specific, well-defined natural language problem with associated data and evaluation metrics [47], designed to test a particular capability of an NLP model (e.g., translation, question answering, sentiment analysis). Specific tasks, benchmark datasets, and metrics are *common knowledge* within the community—for example, every professional NLP researcher is expected to know the field's basic tasks, benchmark datasets, and metrics. This helps in creating a common conceptualization of the field and aligns the researchers' efforts to contribute to well-defined problems with measurable success. These problems are, by *vox populi*, commonly decided to be worthy of pursuit.

In NLP research, contributions are assessed on a specific and well-defined *task* (e.g., machine translation from one language to another, sentiment classification in a given context). The task definition clarifies *the exact problem that the contribution is intended to solve*. There is *some* room in terms of varying the generalizability of the contribution. For example, in NLP, one could argue for a contribution in language translation between *any given* two languages [57] or specifically between two languages (e.g., German and English). In broader terms, similar to NLP formulation of tasks such as text classification, machine translation, and question answering [91], computational tasks for DDP development should provide clear specifications for inputs, outputs, constraints, and evaluation criteria. Such standardization would enable systematic comparison between approaches and facilitate reproducible research in DDP development. Following this logic, in DDP research, one could argue for a contribution, for example, in "user segmentation task" (general) or "user segmentation task when working with tabular data that contains demographic and behavioral information" (specific). Articulating the boundaries of the contribution is therefore essential.

Finally, it is vital to observe that resources can also be combined into *benchmark suites*, which are based on the logic that evaluating a range of tasks provides a more robust measure of general language understanding [91]. Computational techniques also contribute to several stages of DDP development, so performance should be evaluated separately for each task with appropriate metrics, after which an overall performance interpretation of the suggested method(s) should be given. However, most previous work reports only *one* part of the DDP generation process, typically the segmentation part, such as clustering [6], with few articles reporting performance on multiple DDP development tasks [38]. It may be that the increasing participation of ML and NLP researchers in DDP development is influencing the design of DDPs and setting expectations for evaluation rigor. For example, Li et al. [48] propose four pathways toward a more rigorous approach to persona development: (1) building frameworks to determine standard persona attributes, (2) establishing theoretical and empirical groundwork for the design of persona profiles, (3) generating datasets and evaluation standards to test and improve persona creation techniques, and (4) promoting cross-disciplinary collaboration with application-specific assessment of potential risks and biases inherent in persona generation processes. Hardy et al. [28] argue that progress in HCI needs to be shown by "more than marketing", advocating robust benchmarking practices as a way forward. We now proceed to investigating computational resources for DDPs.

Table 1: Computational resources in NLP research and how they can be applicable in DDP research.

Component	What is it?	Why is it needed?
Task	The specific problem the core contribution is intended to solve.	A description of the task (e.g., data merging, data quantification, user segmentation, DDP delivery). The task needs to be specific enough to be solvable.
Datasets	The data used for testing the core contribution on the defined task. Typically, this would be the data for DDP development.	Standardized, well-defined, and clearly documented datasets. If using ML, this typically includes (1) training set, (2) validation set, and (3) test set.
Metrics	The quantitative measure(s) used to assess how well the core contribution performs the task on the test dataset.	Clearly defined, appropriate metrics for the task. Multiple metrics are often needed to capture different performance aspects.
Baselines	Existing, well-understood models, algorithms, or simpler methods used as a point of comparison for the given task.	To contextualize the performance of the core contribution. Is it significantly better than previous state-of-the-art, a standard established method, or a simple heuristic?

3 Development of PersonaBase

In this section, we present the resources in PB, which are available on GitHub[6].The searches for datasets, prompts, systems, and repositories reported in this section were carried out in August 2025. The repository also includes an FAQ that covers prerequisites, intended audiences, and how to get started.

3.1 Design Taxonomy

We created a taxonomy that provides a classification scheme for the design of DDP resources. The taxonomy is coded such that the first two letters indicate the resource type: **PD** for Persona development tasks, **PE** for Persona evaluation tasks, **DS** for Persona datasets, **PP** for Persona prompts, **PS** for Persona systems, and **PR** for Persona repositories (with PD and PE specifically organized under 'Notebooks' in the PB repository). The type is followed by a numerical identifier and a letter suffix denoting the data source. The suffix 'a' signifies that the resource uses simulated or synthetic data, while 'b' indicates it employs real, anonymized data. For example, PD01a would refer to the first persona development task resource that utilizes simulated data, whereas PR03b would denote the third persona evaluation task using real anonymized data. The purpose of this taxonomy is to provide a clear and scalable naming convention for organizing and distinguishing between different DDP resources based on their type and characteristics.

3.2 Tasks

Conceptually, when dealing with DDP tasks, one has to separate two different types of tasks: (1) tasks that persona users would carry out using personas (e.g., designing better products; sometimes called "design tasks") and (2) computational tasks required to develop personas. Here, we refer to the latter tasks by how 'task' is understood in NLP. Research has identified several tasks on the *use* of personas (see a review in [76]), but has not clearly specified (from a computational point of view) the tasks related to the development of DDPs. Again, the issue is not that personas would not have been used for a variety of tasks—they have—but that these tasks are not *formalized* in a way to make them "computational challenges" that could be precisely resolved and measured. There is no inherent objection to giving thought to formalizing the tasks.

As an example, consider this formalization of the "creative design brainstorming task": "*Generate as many novel design ideas as possible for a solar-powered urban trash can that addresses both waste management and renewable energy needs in dense urban environments. Each idea should briefly describe functionality, key features, and potential user benefits.*" It is an example, but the reader can observe the important elements: (a) a clear objective (i.e., "generate as many...") and (b) a precise measure (i.e., the number of design ideas). Of course, the example needs to be further developed, e.g., to consider the quality of the design ideas. But quality could be defined and measured quantitatively, e.g., using expert ratings.

Previous research has taken steps toward computational tasks. For example, Chapman [11] introduced the idea of quantifying personas as information, which has since been carried out in multiple studies [33, 74] as a means of transforming personas into an analyzable format, usually by counting attributes and computing ratio metrics or distributions. This quantification is useful for the evaluation of personas, but the field should define computational tasks more holistically to cover different steps of persona development. Although research has reported variation in persona development steps [34], it has also observed core similarities [41]. Thus, we propose tasks that correspond to methodological steps in DDPs and can be addressed using computational techniques (see Table 2).

To facilitate researchers' navigation around DDP tasks, PB contains 11 notebooks (see Table 3) divided into two types: (1) five persona development notebooks (PD01a–PD05a) and (2) six persona evaluation notebooks (PR01a–PE06b). These use cases (development and evaluation) focus on persona development and evaluation tasks; we leave the persona application use cases for future work. The notebooks were generated using Anthropic's Claude 4-Sonnet, with prompting provided by the lead researcher (Associate Professor), who has multiple years of experience with computational techniques and DDPs. The code and its outputs were first verified by the lead researcher and in case of any inconsistencies (there were some), Claude was asked to provide fixes. Second, the code was verified by a PhD researcher with multiple years of programming and data science experience from industry projects. As such, the code in the notebooks we provide was duly validated.

[6]https://github.com/joolsa/PersonaBase

Table 2: Taxonomy of computational persona tasks. This taxonomy exemplifies how persona generation could be treated as a series of inter-related computational tasks. The resources in PB focus on persona development and evaluation tasks.

Category	Task Name	Description
Persona Development Tasks	Attribute Selection	Identifying relevant user characteristics such as goals, tasks, contexts of use, technical proficiency, and interaction preferences through data analysis to define meaningful persona dimensions for user-centered design.
	User Segmentation	Partitioning users into distinct behavioral groups based on usage patterns, interaction styles, task preferences, and system engagement to create homogeneous segments representing different user archetypes.
	Persona Enrichment	Augmenting persona profiles with contextual details about user goals, pain points, workflow scenarios, accessibility needs, and usage environments to support empathy-building and design decision-making.
Persona Evaluation Tasks	Accuracy Measurement	Validating that personas faithfully represent actual user behaviors, goals, and characteristics by comparing persona attributes against behavioral data, interaction logs, and user research findings.
	Coverage Measurement	Assessing whether a set of personas captures the full diversity of user needs, abilities, contexts, and interaction patterns to ensure no significant user group is overlooked in the design process.
	Diversity Measurement	Evaluating distinctiveness between personas across dimensions relevant to interaction design such as technical expertise, task approaches, assistive technology use, and environmental constraints.
	Perception Measurement	Assessing how designers and stakeholders engage with personas through comprehension studies, empathy evaluations, and analysis of how personas influence design decisions and feature prioritization.
Persona Application Tasks	Persona Filtering	Selecting relevant personas for specific design activities, interface components, or feature development based on task relevance, user journey stages, or interaction contexts.
	Persona Ranking	Prioritizing personas based on design impact criteria such as frequency of interaction patterns, criticality of user needs, accessibility requirements, or alignment with core use cases.
	Persona Updating	Maintaining persona relevance through computational pipelines that integrate new behavioral data, usage analytics, and interaction patterns to reflect evolving user needs and system usage.

This validation covered three criteria: (i) each notebook runs from start to finish in a clean Google Colab environment, (ii) its outputs correspond to the procedure described in the notebook text, and (iii) the metrics follow the definitions in the sources they are taken from. The inconsistencies we corrected included, for example, a notebook returning fewer personas than the number of clusters requested, coverage computed against the sampled data instead of the full baseline, and generated code that did not fit the task at hand.

In terms of content, the persona development notebooks (`PD01a to PD05a`) demonstrate 5 algorithmic approaches to persona generation, including non-negative matrix factorization for YouTube audience behavior (`PD01a`), clustering analysis (`PD02a`), principal component analysis (`PD03a`), Gaussian mixture models (`PD04a`), and LLM-based methods (`PD05a`). Overall, the notebooks demonstrate five different algorithmic approaches to persona generation (PD01a-PD05a) and six evaluation frameworks (PR01a-PE06b) with executable code that researchers can directly replicate and adapt.

3.3 Datasets

In NLP and ML research, performance is measured using standardized datasets. These should be conceptually linked to the task and solvable using the core contribution–that is, there is a logical compatibility between the task and the dataset. In NLP, platforms such as Hugging Face and Kaggle [47] *de facto* institutionalize the principle of sharing datasets. For example, there are currently more than 2600 datasets[7] publicly available in Hugging Face (a model repository for NLP).

In contrast, DDP development often relies on proprietary or task-specific datasets, making direct comparison of methods difficult. Establishing publicly available benchmark datasets would provide a common ground for testing and comparing algorithms, similar to benchmark datasets in NLP [72]. Undoubtedly, NLP/ML datasets are more frequent because these research communities share computational resources more frequently than the HCI community. On the HCI side, the sharing of persona datasets is much rarer, possibly due to privacy, property, and strategic reasons to not disclose real data about users or customers. As an example of NLP-based

[7] https://huggingface.co/docs/datasets/v2.2.1/en/index (accessed on 15 December, 2025)

Table 3: Notebooks in PB. PD = Persona development; PE = Persona evaluation.

ID	Description
PD01a	**NMF.** Demonstrates non-negative matrix factorization to discover patterns in YouTube audience behavior and create personas.
PD02a	**Clustering.** Demonstrates persona creation using clustering analysis.
PD03a	**PCA.** Demonstrates principal component analysis to identify user personas from behavioral data.
PD04a	**GMM.** Demonstrates persona creation using Gaussian mixture model clustering.
PD05a	**LLM.** Demonstrates systematic persona generation using LLMs.
PE01a	**Perception.** Measures perception of generated personas on the persona perception scale.
PE02a	**Diversity.** Measures diversity of generated personas as an evaluation metric.
PE03a	**Coverage.** Analyzes how different persona set sizes affect dataset coverage.
PE04a	**Usage.** Computes and analyzes behavioral metrics for interactive persona systems.
PE05b	**Benchmarking (simple).** Compares three segmentation techniques (NMF, PCA, Clustering) for persona generation.
PE06b	**Benchmarking (advanced).** Creates an evaluation framework combining four metrics (fairness, diversity, coverage, consistency) and compares five generation techniques.

persona datasets publicly available, one frequently cited one is the PersonaChat[8] dataset [95].

However, the NLP/ML datasets are generally *not applicable* to HCI research due to the fundamental reason that HCI and NLP define "persona" differently. In HCI, personas are research-based representations of real user groups used as design and communication tools for humans [17]. In NLP, personas are typically short, fictional textual descriptions (e.g., "I like dogs," "I am a teacher," "My favorite food is pizza") for conditioning or training AI models (often chatbots) [95]. So, these short snippets do not correspond with deep user research for representing specific, validated user segments. A useful separation to understand the difference is that, in HCI, personas help answer "Who are we designing for?", whereas in NLP, personas answer the question, "What identity should this system project?".

To locate DDP datasets, we searched computational repositories (GitHub, Kaggle, HuggingFace) for datasets mentioning personas. We included all the HCI-related datasets we could find and a sample of the NLP-related datasets, as involving all NLP-related datasets would not serve this study's purpose for building HCI-oriented repository for DDP development. We are aware of the fact that our inclusion of NLP/ML datasets mentioning personas is not comprehensive; however, it is adequate for making the analytical reasoning

[8]PersonaChat is a large, crowd-sourced (n = 160,000 utterances from nearly 11,000 dialogues) dataset developed by Facebook AI Research (FAIR) specifically for training and evaluating persona-based conversational AI models.

Table 4: Persona datasets in PB organized by application type. R = real, S = synthetic.

Application	ID	Dataset	Type	Size
Persona Development (user data) (HCI)	DS01	YouTube Audience (2011)	R	1,954
	DS03	Predictive-Personas (2023)	R	2,240
Persona Dialogues (NLP)	DS02	Persona-Chat (2018)	R	8.9K
	DS04	CharacterChat (2023)	S	10K
	DS05	Synth-Persona-Chat (2023)	S	11K
	DS12	AlignX (2025)	S	1.2M
Persona Descriptions (NLP, HCI)	DS07	FinePersonas (2024)	S	21M
	DS08	PERSONA (2024)	S	1.5K
	DS09	Personas-addicted (2024)	S	450
	DS10	Nemotron-Personas (2025)	S	100K
	DS11	PersonaHub (2025)	S	1B
	DS06	Deepfake personas (2023)	S	4

that we do. For a more thorough review of NLP/ML persona datasets, we direct the reader to Chen and colleagues' review of 22 datasets used in personalized dialogue research [15].

As such, PB contains 12 datasets that can be used for developing and evaluating personas, as well as conducting application tasks like updating of persona sets. The metadata we provide about these datasets includes (1) ID, (2) Dataset name, (3) Year released, (4) Description Realness (synthetic or real), (5) Domain type (HCI, NLP/ML, Other), (6) Content type, and (6) URL to dataset.

The datasets in PB suggests a change from real, human-generated persona data (2011-2018) to predominantly synthetic, AI-generated personas from 2023 onward, with 9 of the 12 datasets (75%) released in the most recent two years (2023-2025), suggesting a trend toward synthetic persona generation that is likely influenced by the progress in LLMs. Therefore, LLMs appear to mark a decisive technological transition in persona generation methodology, especially in the provision of novel datasets to experiment with.

Moreover, the datasets are predominantly oriented toward NLP applications (n = 8, 67%) compared to HCI-focused uses (n = 4, 33%), reflecting the focus of the NLP field on conversational AI and language model alignment. For this same reason, many of the datasets actually contain brief text snippets that are presented as "personas", or a set of structured data containing attributes like demographics, hobbies, etc. There are only two datasets (DS01 and DS03) focused on persona development; that is, containing individual or grouped data about people toward the purpose of using that data to develop personas, with the two other HCI-focused datasets developed for analyzing immersive deepfake personas (DS06) and demographic and other biases in LLM-generated personas (DS09). The rest of the datasets contain personas in three types: (1) persona descriptions (n = 5, 42%), (2) persona dialogues (n = 4, 33%), and (3) deepfake personas. One can also observe that the dialogue-focused datasets have been gradually replaced by persona descriptions optimized for LLM alignment.

The datasets also exhibit a clear evolution in scale and complexity, from early small-scale collections with simple biographical

facts (e.g., Persona-Chat's (DS02) 4-5 sentence profiles across 10K dialogues) to recent massive repositories containing millions of multidimensional personas with rich demographic grounding, psychological attributes, and domain-specific expertise (e.g., PersonaHub's 1 billion personas (DS11), Nemotron's 100K personas with 16+ attributes (DS10)). Serving as a good example of HCI-oriented DDP datasets, Hsu et al. [30] performed ML experiments for DDP development *and* shared the data for others to build on. This contribution demonstrates many of the principles seen in NLP and ML domains; for example, the researchers partitioned the data into training (60%) and testing (40%) sets to evaluate the personas.

Additionally, in the datasets, one can observe a strive for diversity, with many of the datasets explicitly mentioning the idea of "being aligned with real populations" (DS08) and representing diverse perspectives (DS07, DS09, DS11). Fourth, while previous persona profiles consisted of approximately 5-15 attributes [64], these computational personas include a clearly higher number of attributes (e.g., 32 (DS10) and 90 (DS12)), which are typically provided as structured data, not as persona profiles. This may be a consequence of the desire to represent diversity better than the handful of personas commonly generated previously. From an HCI perspective, the increase in the number and complexity of personas imposes new interaction challenges because persona users can struggle with cognitive overload related to information-heavy persona sets [42]. To this end, DDP researchers emphasize the need for tools and techniques that make personas more accessible and navigable [36, 78].

Overall, the datasets collected for PB address the gap in shared, standardized data for benchmarking by providing both synthetic and real anonymized datasets that enable researchers to test and compare persona generation methods without requiring proprietary organizational data.

3.4 Metrics

In NLP research, objective assessment requires appropriate *evaluation metrics*, which are quantitative measures chosen for the specific task. Often, a *measurement framework* with multiple metrics is used to capture different aspects of performance. For example, in NLP, BLEU is a single metric specifically designed to evaluate machine translation quality [68], while GLUE [92] contains multiple metrics to assess various aspects of natural language understanding. Again, as with the previous resources, the metrics must be compatible with the task and the dataset.

We reviewed HCI literature to identify metrics used in DDP evaluation (see Table 5 for a summary). Although evaluation is repeatedly mentioned as the weak point of personas [11, 75], we observe that researchers have applied many metrics to determine persona quality and its effects on outcome variables, such as user perceptions and decision making. Interestingly, when personas are provided in interactive persona systems [40], web analytics techniques can be applied to record behavioral metrics about the persona user (see Table 6). In addition to statistics, information theory, or ML metrics, there are some "custom metrics" for DDPs (see Table 5), which often quantify persona sets [11, 74].

Table 5: Types of metrics for evaluating data-driven personas.

Metric Type	Examples
Statistical Validity (Comparing persona to data)	Chi-squared tests, Cosine similarity, Correlation, Set divergence, Accuracy vs. baseline data [6, 11, 53, 85]
User Perception (How users perceive the persona)	Credibility, Consistency, Completeness, Clarity, Empathy, Willingness to Use, etc. [65, 81]
Algorithmic Properties (Characteristics of the generation process)	Stability, Consistency, Diversity, Coverage [74]
User Behavior (How users interact with personas or systems)	Task success rate, completion time, completion steps, recall; eye-tracking (gaze duration, fixations); mouse-tracking (clicks, movements) [78]
ML Model Performance (For underlying ML components)	Accuracy, F1 Score [4, 30]

A challenge is that it is not immediately clear what metrics to optimize in DDP development. DDP research has no clear equivalent for accuracy (or its derivations, like the F1-score). From a computational perspective, a key consideration is that the unit of analysis is often *persona set*, not individual personas. These persona sets are often quantified by their collective persona attributes [11], then the quantified information is compared using metrics like diversity (i.e., the variability of the attributes) [74], with the idea that high-quality personas represent different types of people [24]. Other criteria for "good" personas mentioned in the literature include lack of stereotypical presentation [51], being memorable and realistic [65], and providing useful and relevant information for design [17]. So, there is *some* idea of what constitutes a good DDP, but this idea has not been formalized so that it would be clearly articulated and generally accepted among DDP researchers. Interestingly, there might be a stronger consensus on what a good persona is *not* (stereotypical, irrelevant, abstract, misleading [52]) than what it is. Nonetheless, reaching a degree of consensus on the definition of a good persona is needed to formulate *clear objectives* for DDP development.

Another factor hindering agreement about shared metrics for DDPs is the broad scope of persona application: the use of personas for decision making can range from any number of user-related small and large decisions in an organization, at operational, tactical, and strategic levels, making it difficult to create a comprehensive map of use cases with associated metrics to determine persona quality in a universal sense. Unlike in domains with objectively verifiable outcomes, persona impact is difficult to define, observe, and measure. For example, in NLP, if we apply a new model to a translation problem (translating text from Language A to Language B) and it works better, we know that the impact is better translation quality. However, when we apply DDPs in decision making, how do we know that the decisions are now "better" for the end users?

These challenges culminate in the difficulty of *establishing ground truth* for persona assessment. Instead of having one (and only one)

[A]

Demographic Group	Video A	Video B	Video C	Video D	Video E
US_Male_18-24	4,567	6,234	8,901	2,345	5,678
US_Male_25-34	7,890	12,456	15,234	4,567	9,876
US_Female_18-24	5,432	7,890	11,234	3,456	6,789
US_Female_25-34	6,789	11,234	13,567	4,321	8,901
CA_Male_18-24	1,234	2,108	3,456	987	2,345
CA_Female_25-34	2,345	3,567	4,789	1,456	3,210
GB_Male_25-34	1,567	2,890	3,678	1,234	2,567
GB_Female_18-24	1,890	3,234	4,123	1,567	2,890
DE_Male_35-44	987	1,678	2,234	789	1,456
JP_Female_25-34	1,123	2,345	2,987	1,089	1,789

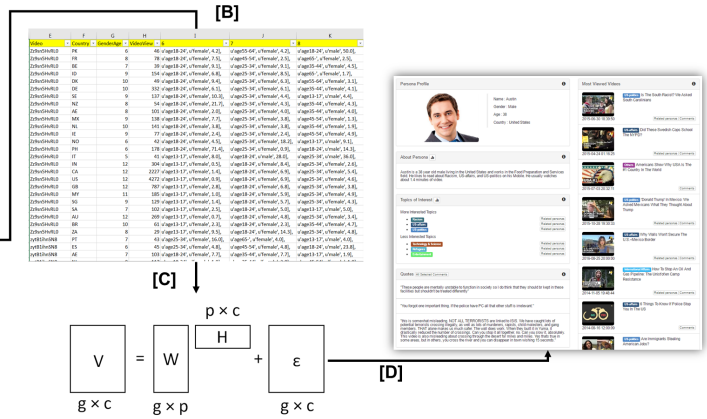


Figure 3: Example of grouped data structure in YouTube Analytics API export [A], belonging to a larger matrix [B]. Each row represents a specific country-gender-age segment, demonstrating the typical sparse data structure in online audience statistics. This matrix can be segmented, e.g., using NMF [C] to skeletal personas that are then enriched to generate full persona profiles [D]. The notebook `PD01a` in `PB` demonstrates how to generate personas from this grouped data structure, whereas notebooks `PD02a`, `PD03a` demonstrate how to create personas from individual-level user data, in which each row represents an individual user instead of a group.

Table 6: Examples of behavioral metrics in interactive persona systems [78]. The `PE04a` notebook in `PB` illustrates how to implement, compute, and analyze these metrics, creating visualizations about user engagement and persona performance.

Persona-Based Metrics	
Time Spent per Persona	Total duration of user interaction with a given persona; a proxy for interest, as longer interaction may indicate higher engagement or confusion.
Number of Visits per Persona	How often a persona is accessed; another proxy for interest, e.g., frequently visited personas are seen as more relevant or engaging.
User-Based Metrics	
Number of Personas Visited	Count of distinct personas accessed by a user in a session; gauges how extensively a user explored the available personas.
Persona Coverage	Ratio of personas a user visited relative to those available; reflects the user's interest in the whole persona set.
Persona Bi- and Trigrams	Frequency with which users visit pairs or triplets of personas in a session; indicates comparative behavior across personas.

set of good personas generated from a given dataset, it is possible to generate many persona sets that are all good, but in different ways [9]. The logical conclusion from this is that *persona evaluation requires the use of multiple metrics simultaneously*, which is what we illustrate with `PB` notebooks. In `PB`, the evaluation notebooks demonstrate multiple ways to assess persona quality by computing and analyzing metrics, including measuring user perception on the persona perception scale (`PE01a`), quantifying diversity within persona sets (`PE02a`), analyzing the relationship between persona set size and dataset coverage (`PE03a`), and tracking behavioral metrics in interactive persona systems (`PE04a`). The evaluation notebooks culminate in comparative analyses that benchmark different segmentation techniques (`PE05b`) and a unified evaluation approach (`PE06b`) that combines fairness, diversity, coverage, and consistency metrics to compare five persona generation methods across varying set sizes, demonstrating how researchers and practitioners can select persona creation strategies based on empirical evidence.

3.5 Baselines

3.5.1 On Baselines in DDP Research. To contextualize the results of a specific algorithm, model, or system that researchers introduce, the performance of this contribution is compared against relevant *baselines*, typically existing state-of-the-art or well-established standard methods that represent known norms in the field. Overall, DDP research has used various methods as points of comparison in their studies (see Table 7).

Baseline choice in DDP development depends on (a) the task being resolved and (b) the data input. This choice could be made based on the popularity of methods in the previous literature. For example, for data organized in matrix structure, the most common segmentation technique in DDPs has been non-negative matrix factorization (NMF) (see, e.g., [4, 41]). This implies that NMF would be the "baseline to beat" when the task is persona segmentation, and the data type is matrix. Similar connections could be made by systematically analyzing different types of datasets and algorithms deployed for them, which we leave for other work.

Research compares personas to specific baselines. For example, personas have been compared against nameless, faceless target groups used in advertising [79], indicating that *non-personified segments* can be a baseline. Another study used personas against

Table 7: Taxonomy of baseline types for persona evaluation.

Category	Baseline Type & Examples
System	Commercial Systems: Analytics platforms [78] Vanilla LLMs: Unmodified LLMs like ChatGPT [27] Alternative UI: Chat interfaces, deepfake personas vs. templates [42, 43, 64]
Method	Manual Creation: Human-generated personas [96] SOTA Algorithms: Segmentation algorithms [4, 41] Non-personified Segments: Nameless target groups [79] Alternative UCD: User stories, JTBD, requirement specs (N/A)
None	Studies without comparative evaluation [41]

YouTube Analytics [77], indicating that *commercial analytics systems* can be a baseline. Different persona templates could be used as competing "versions" of the persona user interface [64], and the decision to use a template or another interaction modality, such as dialogue or chat [42], is another opportunity for baseline, with the classic "persona profile" [64] serving as the standard baseline for persona design. Testing new modalities, such as chat and deepfake personas afforded by GenAI technologies [42, 43] is particularly pressing because these modalities challenge the 20-year-old tradition of providing personas as "templates," "profiles," or "narratives" [17, 61, 64]. For LLM-generated personas, standard LLMs such as ChatGPT provide a baseline [27], along with comparing AI-generated personas with manual persona creation [96].

We specifically focus on three types of resources that make sense for computational benchmarking of DDPs: (1) algorithms, (2) prompts, and (3) systems. These resources are compatible with DDP development practices: algorithms are often used for segmentation in the DDP process; systems could be compared by their ability to generate personas; and prompts are increasingly driving persona generation, so comparing different prompts forms natural baselines.

3.5.2 Algorithms. Algorithms are commonly applied in DDP development to segment a given user dataset into "skeletal personas" [98] that are then enriched using other computational techniques (e.g., topic modeling [4]) to produce full persona profiles [38, 41]. Algorithms that have been used for segmentation in DDP development include at least K-Means [46] and other clustering techniques [6], latent semantic analysis (LSA) [55], Latent Dirichlet Allocation (LDA) [86], non-negative matrix factorization (NMF) [4], and principal component analysis (PCA) [84]. As resources, these algorithms are based on open source contributions and the software implementing them, such as `scikit-learn` [69].

Research shows that DDP developers typically use "off the shelf" options instead of developing their own tailored persona segmentation algorithms. A major concern in using preexisting algorithms from other fields, such as ML, is that these algorithms were not designed primarily for user segmentation, but for other purposes, such as dimensionality reduction for more efficient statistical learning [59]. Consequently, DDP researchers have expressed concern about the negative side effects of using standard data science algorithms for persona segmentation, such as their emphasis on central tendencies and majority trends [12] that risks omitting minority segments, which would be of interest in many HCI applications [23]. Similar problems of "excessive averaging" exist when using pre-trained models that represent general Internet biases in LLM-generated personas [32, 80].

However, despite the reliance on "out-of-domain" algorithms, DDP research also demonstrates some contributions to algorithmic development, or at least their application in the persona domain. As a "textbook example" of how to compare algorithms in persona generation, Brickey et al. [6] showcase an experimental framework that uses expert manual clustering as the gold standard. The researchers systematically evaluated three semi-automated methods for persona development (LSA, PCA, and cluster analysis) on the same data set, and used Cohen's kappa to measure agreement with human-generated clusters. In a similar vein, Ditton et al. [21] developed a semi-automated hyperparameter tuning framework for clustering algorithms in persona development. The researchers applied Average Feature Significance (AFS) designed to reward clusters with statistically significant differences between centroids, and demonstrated their framework can automatically eliminate 40% of algorithm-parameter combinations while producing personas that align with established behavioral theory. In addition to comparing several clustering algorithms for persona development, the work illustrates algorithmic development to support hyperparameter setting in persona development.

In turn, Salminen et al. [74] compared six algorithms for persona generation, applying diversity, fairness, and consistency to assess how algorithms shape the demographic representation of DDPs, finding that the choice of algorithm fundamentally impacts the attributes of the generated personas.

3.5.3 Prompts. A rising genre of computational resources in DDP development is prompts; i.e., instructions for LLMs to generate personas or assist in specific DDP development tasks. The interaction with many generative models, especially LLMs, relies on effective prompting [94]. Researchers have shared prompts for persona generation. Foundational prompts focus on *initial persona creation*, instructing LLMs to generate descriptions based on specified demographics, user data [82], or predefined characteristics [14]. Other prompts focus on *persona refinement and expansion*, modifying or elaborating on existing personas by adding attributes or context [13]. Furthermore, personas are frequently employed in *downstream tasks*, with prompts instructing LLMs to act *as* a persona or use its traits to generate outputs for simulations, evaluations, or content creation [5, 88]. Finally, prompts are designed to *evaluate* the quality of the generated personas or their outputs using structured criteria or rating scales [67, 73].

Persona research can benefit from the systematic collection of prompts for DDP development, along with a meta-analysis of how different prompting strategies affect outcome variables. This is because current prompting approaches for DDP development vary by structure, specificity, and theoretical foundation, leading to inconsistent results [3]. This variability is particularly problematic when prompts lack explicit grounding in persona theory or fail to incorporate domain-specific contextual requirements. Templates and guidelines for structuring prompts could improve the consistency of the developed DDPs. Therefore, we include LLM prompts as part of the PB. For this, we extensively reviewed articles published on

Table 8: Persona prompts in PB organized by the primary focus of the prompting.

Purpose	Frequency
Persona Generation Methods	**n = 20, 74.1%**
Direct generation prompts	n = 10, 37.0%
Iterative/workflow-based generation	n = 4, 14.8%
Domain-specific generation	n = 6, 22.2%
Persona Evaluation & Application	**n = 7, 25.9%**
Evaluation frameworks	n = 3, 11.1%
Interactive systems and tools	n = 3, 11.1%
Applied use cases	n = 1, 3.7%

HCI top venues (e.g., ACM CHI, IUI, DIS) since the introduction of LLMs, and we also include *arXiv* publications in this search because some of the articles reporting persona prompts were only available as preprints at the time of carrying out the investigation. The systematic procedure for these searches, which were conducted until August 2025, and the associated screening is reported in related work [3]. We located 27 articles that share their prompts for persona development (see Table 8).

We divided the prompts manually into two main categories: (1) persona generation methods and (2) persona evaluation and application. Prompts in the *Persona Generation Methods* category (n = 20, 74.1%) focus on creating personas. This includes direct generation prompts using skeletal creation followed by expansion (e.g., PP02), demographic-driven synthesis (e.g., PP05), attribute-based generation (e.g., PP10), and two-stage summary-to-detailed approaches (e.g., PP27). The category also includes iterative and workflow-based methods such as multi-stage processes combining interview analysis with trait extraction (e.g., PP11), rule-based validation workflows (e.g., PP21), and thematic analysis pipelines (e.g., PP25). Domain-specific approaches are represented as well, including character imagination for creative writing (e.g., PP03), review-based persona development (e.g., PP06), personality reconstruction from descriptions (e.g., PP13), DDP steering (e.g., PP16), and UX-focused generation (e.g., PP22). The *Persona Evaluation and Application* category (n = 7, 25.9%) focuses on using, evaluating, or interacting with already-generated personas. This includes evaluation frameworks for coherence and consistency rating (e.g., PP04), quality assessment metrics (e.g., PP12), and Likert-scale validation of users' perceptions of the personas (e.g., PP14). Interactive systems are also represented, such as audience understanding tools with chat-based persona interaction (e.g., PP18), emotion-driven conversational personas (e.g., PP23), and writer-defined feedback generation (e.g., PP24). Applied use cases include, for example, persona variants for content recommendation auditing (e.g., PP01). These results suggest that most research effort has focused on persona generation methodologies, while a smaller portion addresses the deployment and validation of personas in practical contexts.

3.5.4 Systems. In general, DDP systems leverage quantitative user data and algorithmic methods to create personas that aim to combine the empathetic benefits of personas with the analytical rigor and scalability of data science [35]. There are both (1) research-driven and (2) industry-driven systems. In our review, we focus on the former. We reviewed past DDP research and the current state-of-the-art articles from premier HCI venues, including CHI, IUI, UIST, and DIS, to identify articles developing persona systems. Through this process, we identified a sample of 19 systems (see Table 9). Note that this list may not cover all existing systems, but it is meant to provide a reasonable sample of persona systems published in HCI venues. In the PB repository, we recorded the following information about each system: (1) ID, (2) Name of system, (3) Year introduced, (4) Description, and (5) URL. These systems can be used as baselines for comparison or as inspiration for further DDP development.

Table 9: Persona generation systems listed in PB organized by technical approach.

ID	System (Year)	Domain
Statistical/ML-Based Systems		
PS01	APG (2018)	Social Media
PS02	Survey2Persona (2022)	Survey Analysis
LLM-Based Persona Generation		
PS03	PersonaGen (2023)	Software Dev
PS06	Crafter (2024)	Software Eng
PS12	LLM-summarized (2024)	UX Design
PS17	Persona-L (2025)	Accessibility
Multimodal Persona Systems		
PS05	Vivid-Persona (2024)	Interactive
PS07	Character-Centric (2024)	Storytelling
PS11	Auto-Generated (2024)	Education
Interactive Agent Simulations		
PS04	Generative Agents (2023)	Simulation
PS08	Agentic Society (2024)	Population
PS15	Gen. Agent Sims (2024)	Behavior
Specialized Application Tools		
PS09	IQA-EVAL (2024)	Evaluation
PS10	Privacy Auditing (2024)	Privacy
PS13	CloChat (2024)	Customization
PS14	Impressona (2024)	Writing
PS16	Proxona (2025)	Interaction
PS18	PersonaFlow (2025)	Research
PS19	SimTube (2025)	Content Creation

An example is the APG system that connects to data sources (e.g., Google Analytics, Facebook Insights, CRM data) through application programming interfaces (APIs) [41]. APG uses algorithms (e.g., NMF) and other data science techniques (Python libraries such as `Pandas, scikit-learn`) to segment users based on behavior and demographics [40], enrich these segments with information, and generate persona profiles provided to persona users through an interactive web interface, including granting access to the underlying user data used for their development [36]. Finally, systems like APG can periodically update the personas [35].

Investigating these systems, we can observe some trends. First, there is a distinct acceleration in development since 2023, with only

Table 10: Persona research repositories listed in PB organized by resource availability. ✓ denotes the resource is available; – denotes it is not.

ID	Repository (Year)	Code	Data	Paper
Documentation & Templates				
PR01	CTS-Personas (2019)	–	–	✓
PR02	AI Personas (2021)	–	–	✓
Complete Research Packages				
PR03	Persona-Generation (2021)	✓	✓	✓
PR06	Predictive-Personas (2023)	✓	✓	✓
PR07	persona-generation-workflow (2024)	✓	✓	✓
PR08	identifying-privacy-personas (2024)	✓	✓	✓
Code & Methodology Tools				
PR04	HyPersona (2022)	✓	–	✓
PR05	Epiphany (2023)	✓	–	–

two systems (11%) introduced before 2023 compared to 17 systems (89%) emerging from 2023 onward. This is likely associated with the widespread availability of LLMs, which has facilitated building systems for user representation in general. Second, the systems exhibit a clear evolutionary trajectory from early specialized data-processing tools handling specific input types–such as APG's PS01 social media analytics (2018) and S2P's PS02 survey data focus (2022)–to the post-2023 explosion of LLM-powered frameworks that span interactive simulations (e.g., PS04), multimodal persona creation combining text, image, and video (PS05), and collaborative workflows (PS13, PS14, PS16). Third, about a quarter of the systems (n = 5, 26%) make their source code (or parts of it) publicly available (PS04, PS08, PS10, PS12, and PS15). The shared source code facilitates the replication of research studies and provides starting points for future system development.

Overall, as our contribution, PB lists, describes, and links to 19 DDP systems, which is useful because it decreases the fragmented systems landscape in DDPs. The catalog of systems enables researchers to learn from, benchmark against, and build on existing implementations.

3.6 Software Repositories

In the process of building PB, we systematically identified other computational resources in GitHub concerning DDPs by searching with relevant keywords ("personas", "data-driven personas"). The vast majority of the repositories we found were incomplete, e.g., with missing README files, missing data for reproduction, etc. Our purpose here was not to cover all possible repositories, so we focused on including *example repositories* that we found containing reasonably well crafted documentation and resource bases. We decided to list these, along with their metadata, including (1) ID, (2) Repository name, (3) Year, (4) Description, (5) Contains code, (6) Contains data, (7) Associated with a research paper, and (8) URL. These repositories (see Table 10) illustrate how to set up computational repositories for DDP projects, thereby serving as examples of reasonable well executed DDP repositories.

Inspecting the repositories, we can observe a progressive maturation from 2019 to 2024, evolving from documentation-focused efforts (PR01 providing descriptive role profiles) and foundational resources (PR02 survey templates and image libraries in 2021) toward increasingly sophisticated computational approaches, including algorithmic comparison frameworks (PR03, PR04), AI-automated generation tools (PR05, PR06), and recent advanced integrations of LLMs with human-AI collaborative workflows (PR07) and specialized analytical methods (PR08). There is a notable trend showing that 75% (6 of 8) of repositories include executable code, 63% (5 of 8) provide datasets, and 75% are associated with research publications, reflecting sound practices in providing reproducible, data-driven, and academically rigorous persona development methodologies. Overall, the repositories address the gap in centralized, quality-controlled resources by aggregating computational materials (code, datasets, and research outputs) from fragmented sources into a collection that demonstrates the content of computational resource repositories. PB differs from the repositories by providing more resources, both in terms of type and quantity.

4 Formative Evaluation Studies

We conducted two formative evaluation studies: (1) an exploratory walk-through study with four persona researchers, and (2) an online questionnaire study among 25 designers and developers experienced with personas, with details reported as follows.

4.1 Walk-Through Study with Persona Researchers

4.1.1 Approach. To evaluate PB, we organized semi-structured Zoom sessions to interview four persona researchers with different levels of experience with DDPs and computational resources (see Table 11). The frame for the semistructured interviews was built in a way that we assessed the participants' (a) Background and Expertise, (b) Feedback on PB, and (c) Additional Insights on Computational Resources. The detailed frame for the walkthrough sessions and semistructured interviews is available in the supplementary material[9].

Two researchers (R01 and R03) can be considered as domain experts in DDPs specifically (corresponding to the "expert persona researcher" we had defined) while the two others (R02 and R04) are less versed in DDPs and more geared toward qualitative design personas (corresponding to the "novice persona researcher" we had defined). All had computer science or HCI degrees, all had both industry and research experience, and all currently work primarily in academic research. The identification of the persona researchers was done based on published research on personas, and they were recruited by direct messages asking for their feedback on an on-going effort to develop computational resources for persona development. We first asked the researchers about their persona experience and computational resource experience; the sessions were organized individually and each researcher was first (1) asked general questions about their background and opinions about DDP resources, then (2) walked through PB by the moderator sharing their screen, and (3) and asked for their feedback about PB, especially on how it could be further improved.

[9] https://osf.io/vn9tu/?view_only=4315c176d4f047e5a32b473602807049

Table 11: Participant demographics and expertise.

ID	Nationality	Age	Gender	Description
R01	South Korea	30s	F	A computer science PhD student with industry collaboration experience who researches persona methods and has developed a persona system.
R02	China	20s	F	A former UX designer turned researcher exploring dynamic personas and data visualization, with interests in computational approaches to persona systems.
R03	United States	60s	M	An academic persona researcher grounded in classical statistics who bridges qualitative persona research with computational methods and industry practice.
R04	Turkey	30s	M	An HCI researcher focused on design methods and ideation who critically examines the intersection of qualitative personas and computational data-driven approaches.

The sessions lasted about half an hour each and they were moderated by the lead author (who has about 8 years of experience in DDP research). The sessions were recorded with the researchers' permission and transcribed for analysis. The moderator took notes during the discussion and analyzed the transcripts with the help of Claude 4.5-Sonnet in a computer-assisted qualitative analysis process similar to Wen et al. [93] that included asking Claude to summarize the sessions and then verifying the summaries by reading them line-by-line and editing for facts and narrative coherence. The results of this inquiry are presented as follows.

4.1.2 Results. The themes from the researcher feedback on PB are shown in Table 12. Inspecting these themes at a higher level of abstraction, we can identify that Navigation/Discovery (T1), Onboarding/Learning Pathways (T3), and Template-Based Resources (T4) seem to address the same underlying problem: *how do users figure out what to do with this repository?* T1 is about finding the right resource, T3 is about understanding the expected workflow, and T4 is about providing scaffolding for action. We could group these themes into a broader theme of *Onboarding Challenges*, which deals with the issue of making sense of the content in computational repositories.

In contrast, Documentation/Purpose Mapping (T2) and Accessibility/Disciplinary Bridging (T6) concern making the repository legible across diverse user groups, which we can refer to as *Cross-Community Translation* and which resembles cross-disciplinary efforts in using the same computational resources (cf. [54]). T2 is about explaining why and when to use something (bridging from tool to purpose), while T6 is about bridging disciplinary paradigms and preventing misappropriation of concepts. Both are about making implicit knowledge explicit for different user groups of PB.

Finally, Competition/Community-Building (T5) is less about the design of PB and more about creating external incentives that motivate engagement and establish shared standards for DDP resources, which is why we can refer to it as *Traction*, which resembles the *cold start problem* of making a new product more popular [87].

So, overall, participants identified needs around (1) orientation within PB, (2) its documentation for different communities, and (3) external mechanisms for sustained engagement.

Based on the researcher feedback, we created a "Persona leaderboard" notebook (`PE06b`, see Figure 4) that implements a standardized evaluation framework for comparing five persona generation algorithms (PCA, NMF, Clustering, Spectral Embedding, and UMAP) using four metrics, namely diversity, fairness, consistency, and coverage. All metrics are normalized to a 0-100 scale where higher scores are better, enabling direct comparison and aggregation. The notebook generates personas of different sizes (k = 5, 10, 15), evaluates each algorithm's performance, creates comprehensive visualizations and leaderboards, and exports detailed results, including an overall champion determination based on averaged scores across all metrics.

We next turn to the evaluation study among persona practitioners.

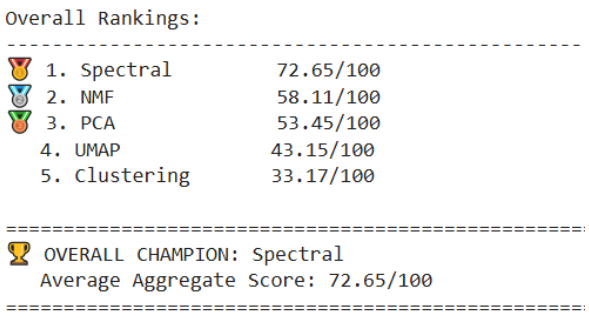


Figure 4: Leaderboard example from `PersonaBase (PE06b)` that we developed following the formative evaluation study. This notebook tests five persona generation algorithms for four evaluation criteria: accuracy, consistency, diversity, and coverage. These results indicate that (a) there are substantial differences in the performance of different persona generation algorithms, and (b) there is room for further algorithmic development because even the best algorithm tested here produces a performance score below 80% of the maximum.

4.2 Online Questionnaire Study

4.2.1 Questionnaire Design. The survey was developed in Qualtrics and consisted of three sections (all survey items are available in online supplementary material on Open Science Foundation[10]). The first section collected demographic information (age, gender, nationality, job role) and assessed the practitioners' self-reported

[10]https://osf.io/vn9tu/?view_only=4315c176d4f047e5a32b473602807049

Table 12: Thematic analysis of participant suggestions for improving PB.

ID	Theme Name	Supporting Quotes
T1	Navigation and Discovery Mechanisms	**R02:** "I'm wondering, could the repository also have this kind of searching feature so the user can come in and even their intent is very vague. They can know (...) are they looking for a data set or are they looking for a handbook, notebook, or actually they are looking for a project." **R02:** "But I, as a researcher, maybe I just need an entry kind of interface or kind of like search interface sort of things. Help me to like explore around the landscape of this maybe a little bit visually and also make sure my intent would be supported by this kind of data."
T2	Documentation and Purpose Mapping	**R01:** "So, actually, the (...) data set part is the most, eye-catching part, I guess, because, [when learning about DDPs] it was very hard to imagine about what kind of persona outcome there should be." **R02:** "[Each notebook should include] explanatory materials addressing why the approach is useful, when researchers should employ it, and what theoretical or methodological foundations underpin it."
T3	Onboarding and Learning Pathways	**R04:** "What is the pipeline [of using the repository]? (...) Where do I start? And what should I expect at the end?" **R04:** "What am I assumed to know? If you never use Python, don't bother?"
T4	Template-Based Resources	**R01:** "[Implement] a block-building interface where researchers could interactively construct prompts by selecting components, with each component traceable to specific papers."
T5	Competition and Community-Building	**R03:** "Doing something like the Semaval folks did might be a good approach, (...) running workshops and competitions in these computational persona areas to get persona folks to use it." **R03:** "There's, you see these type of benchmarks and competitions in different areas. Again, to my knowledge, it hasn't been done in personas because, you know, you'd have to do a lot of stuff in personas also in terms of setting the context."
T6	Accessibility and Disciplinary Bridging	**R03:** "I think it is definitely set up for the computational persona researchers. A lot of the, a lot of persona researchers will not be able to access it. It is beyond their capability and not really even in their paradigm to do this kind of thing." **R04:** "So the persona is just their content to align that chatbot, for example, to speak differently when it assumes the role of different persona... Yeah, that is, you know, concept abuse."

expertise levels regarding research orientation, personas, DDPs, and computational resources. The practitioners also rated the relevance of seven types of DDP resources to their professional goals and ranked these resources by perceived usefulness. In the second section, administered after the practitioners browsed the PB GitHub repository for approximately five minutes, they provided open-ended feedback on the repository's strengths and weaknesses, rated their agreement with five statements about the repository's clarity, organization, and utility using five-point Likert scales, and assessed the usefulness, clarity, and completeness of five resource categories (datasets, notebooks, prompts, systems, and repositories). The survey concluded with open-ended questions asking the practitioners to elaborate on their responses and describe the target users of the repository.

4.2.2 Participant Recruitment and Data Processing. Within the scope of our limited data collection budget, we recruited 27 practitioners through Prolific, a crowdsourcing platform for online research [66]. Participants were prescreened based on their current job role (including UX/UI designers, product designers, software engineers, AI/ML engineers, and web developers) and industry (including software, consulting, market research, and scientific or technical services) to ensure relevance to our study on DDPs. Eligibility was further restricted to individuals who had used personas in their professional work. The study was administered via Qualtrics on desktop devices, with an estimated completion time of 15 minutes (5 minutes browsing a GitHub repository and 10 minutes completing a survey). Participants were compensated £3.00 (equivalent to £12.00/hr, which was labeled as "Great!" compensation by Prolific's calculator). To ensure data quality, we enabled Prolific's authenticity checks for detecting suspicious behavior and automatic rejection of exceptionally fast submissions. However, we also manually reviewed the data before the analysis. For that, we downloaded the data from Qualtrics and cleaned it by removing one pilot tester, who is not counted among the 27 participants, one participant who did not provide sensible answers to open-ended questions, and one participant whose open-ended responses appeared to be AI-generated. Although Prolific's authenticity check flagged four responses as 'mixed,' all responses were manually reviewed by the lead researcher regardless of the automated flags. After cleaning, 25 (92.6%) valid responses remained for analysis, resulting in a rejection rate of 7.4%, which is aligned with previous work showing high-quality responses from Prolific [22].

4.2.3 Practitioner Information. The practitioners had a mean age of 29 years (SD = 7.19) and were predominantly male (n = 18, 72%). The practitioners represented diverse nationalities from Africa, Asia, Europe, and the Americas, with the largest representations from Kenya, Portugal, India, and South Africa. Most practitioners held technical roles, primarily software engineers and developers, with additional roles including data analysts, product designers, and managers. Regarding research orientation, the majority identified as mixed-method specialists (n = 16, 64%), followed by qualitative (n = 5, 20%) and quantitative (n = 4, 16%) specializations. Also, most practitioners reported advanced expertise in persona use ("I have used in my job several times"; n = 18, 72%) and computational

resources (n = 17, 68%), while about half (n = 13, 52%) reported advanced expertise specifically in DDPs.

4.2.4 Results. **Resource relevance:** Prior to browsing PB, the practitioners rated the relevance of seven types of DDP resources *in general*, not specific to PB, which was evaluated after Part 1 of the study. This was done by asking the practitioners to rate the relevance of different resources to their professional goals on a five-point Likert scale (see Figure 5). Datasets were rated most highly, with 44% of the practitioners finding them extremely relevant and 72% rating them as very or extremely relevant. Software repositories and LLM prompts were also perceived as highly relevant, with 80% and 72% of practitioners rating them as very or extremely relevant, respectively. The systems database (76%) and evaluation notebooks (68%) similarly received favorable ratings. Algorithm notebooks were rated as very or extremely relevant by 68% of the practitioners, though they also received the highest proportion of negative ratings (8% rated them as slightly or not at all relevant). Leaderboards were perceived as least relevant overall, with only 60% of the practitioners rating them as very or extremely relevant and 16% rating them as slightly relevant. This was surprising, but might stem from the fact that practitioners are likely less interested in research benchmarking, which leaderboards represent, than actionable resources like code and datasets.

Assessment of PersonaBase: After providing their general feedback on the relevance of computational resources for their work, the practitioners were provided a link to browse the PB repository on GitHub for approximately 5 minutes. They were asked to read the information on the homepage of the repository, visit its folders, and open the resources, including notebooks, to see what they are about. After this, the practitioners were asked to return to the survey and answer questions about PB.

Analyzing the results, we can see that, overall, the practitioners assessed the coverage of PB's computational resources positively. The majority of the practitioners (64.0%, $n = 16$) indicated that the repository covers all the different computational resources well, while approximately one-third (36.0%, $n = 9$) felt that the repository covers different resources reasonably well but something is missing. Notably, no practitioner selected the option indicating that the repository does not cover computational resources well, suggesting that PB meets the fundamental expectations for resource coverage in DDP development.

We also asked the practitioners for their general feedback about PB, as well as their perceptions of PB's usefulness, clarity, and completeness, again, using a five-point Likert scale (see the online supplementary material on Open Science Foundation[11] for the measurement items). Concerning the practitioners' general feedback about PB, the results in Figure 6a indicate a positive assessment. The practitioners found the purpose of the repository clear ($M = 4.48$, $SD = 0.59$) and considered the resources well organized ($M = 4.48$, $SD = 0.77$). Importantly, the practitioners disagreed that the repository was missing key resources ($M = 2.96$, $SD = 1.17$) or contained too many items ($M = 2.40$, $SD = 1.00$), suggesting that `PersonaBase's` scope and coverage are reasonable. Furthermore, the practitioners expressed a strong intention to use the repository for DDP related tasks ($M = 4.32$, $SD = 0.63$). Concerning the perceived usefulness, clarity, and completeness of PB resources, the practitioners rated all resource types favorably (see Figure 6b), with most means fall between 3.8 and 4.5 on a 5-point scale. Looking at these ratings, the differences are quite small, which suggests a generally favorable perception.

Differences by prior experience: To investigate whether previous persona or DDP experience influenced the practitioners' ratings on the above variables, we conducted a series of Mann-Whitney U tests comparing two groups: *Advanced* ($n = 18$ for persona experience and $n = 13$ for DDP experience) and *Intermediate/Novice* users ($n = 7$ for personas and $n = 12$ for DDPs). For this comparison, we grouped the Novice and Intermediate practitioners into group, due to their lower frequency, and compared this group with the Advanced practitioners. When comparing groups by different levels of persona experience, only one significant difference emerged: advanced persona users rated the *completeness of datasets* higher ($M = 4.33$) than Intermediate/Novice persona users ($M = 3.43$), $U = 97.0$, $p = .031$, $r = -.54$. The rest of the differences were not significant.

However, when comparing practitioners with less and more DDP experience, the results indicated significant differences on four variables (see Figure 7). First, Advanced DDP users rated *the relevance of prompts for generating personas using LLMs* significantly higher ($M = 4.46$, $SD = 0.66$) than Intermediate/Novice users ($M = 3.50$, $SD = 1.05$), $U = 126.5$, $p = .006$, $r = -.62$. Second, Advanced DDP users rated the *relevance of leaderboards comparing DDP methods* higher ($M = 4.08$, $SD = 0.86$) than Intermediate/Novice users ($M = 3.08$, $SD = 1.16$), $U = 122.5$, $p = .011$, $r = -.57$. So, this indicates that persona generation prompts and leaderboards are perceived particularly important by practitioners with more DDP experience.

Third, Advanced users rated *dataset completeness* significantly higher ($M = 4.46$, $SD = 0.52$) than Intermediate/Novice users ($M = 3.67$, $SD = 0.82$), $U = 117.5$, $p = .024$, $r = -.51$. Interestingly, Advanced users also agreed more with the statement about the *repository containing too many items* ($M = 2.85$, $SD = 1.14$) than Intermediate/Novice users ($M = 1.92$, $SD = 0.90$), $U = 118.0$, $p = .025$, $r = -.51$. This may suggest that advanced DDP users are more focused on certain resources, especially datasets, and consider a large selection of resources more cumbersome than newcomers, who in turn may benefit more from a broader set of resources to learn about DDP development.

Open-ended insights: Finally, to gain more in-depth insights, we also asked the practitioners open-ended questions about the strengths, weaknesses, and future development suggestions for PB (again, the questions are available in the Open Science Foundation repository). When analyzing this feedback, we applied the common academic logic of positioning current findings to previous ones, namely, that the findings can *corroborate*, *contradict*, or *complement* prior findings. By "prior findings", here we mean the results of the thematic analysis of persona researcher walkthrough sessions (see Table 12).

Overall, the practitioner feedback largely corroborates the researcher interview themes but also contained some additional insights. In terms of *corroboration*, the practitioner feedback most strongly validates Documentation and Purpose Mapping (T2) and

[11] https://osf.io/vn9tu/?view_only=4315c176d4f047e5a32b473602807049

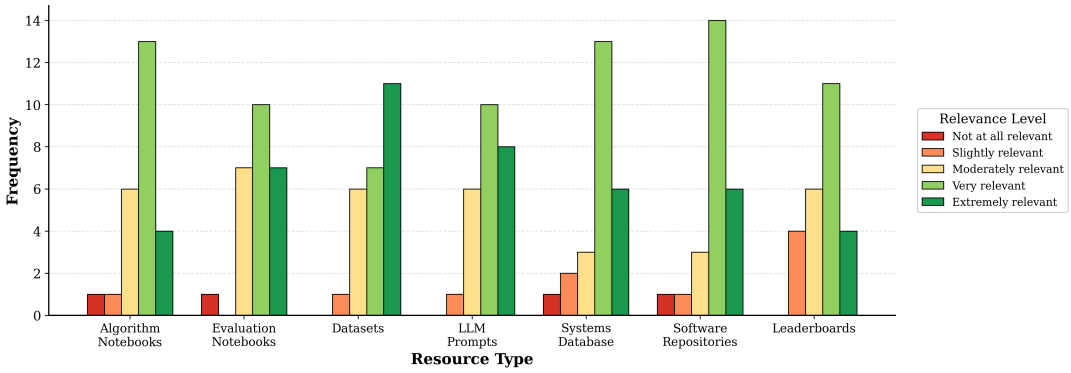
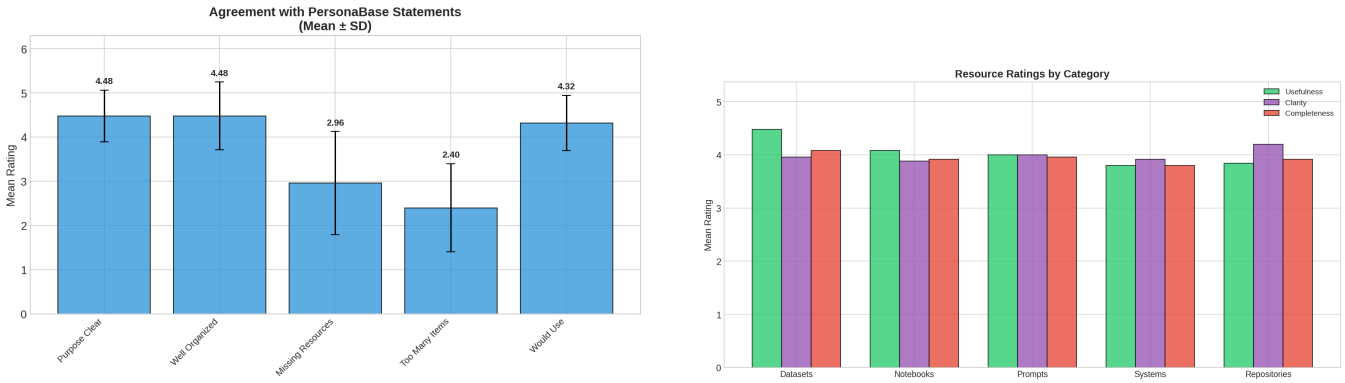


Figure 5: Practitioners' assessment of DDP resource relevance in general.



(a) **(b)**

Figure 6: Overview of practitioner evaluations on data-driven persona resources.

Onboarding and Learning Pathways (T3). The practitioners' requests for "more detailed documentation" (P06), "simple introductory guides" (P10), "step-by-step videos" (P18), and "end-to-end workflows" (P06) are similar to the researchers' requests for explicit documentation of purpose and use cases, as well as comprehensive orientation guides and workflow demonstrations. The concern that PB is "a bag of resources... instead of a single reproducible pipeline" (P15) exemplifies especially advanced practitioners' suggestions for clear pipelines with defined starting points and expected outcomes. Competition and Community-Building (T5) is also corroborated, with practitioners noting the repository is "still new and not widely used" (P07) and requesting "contribution guides and contact information" (P04), pointing to a similar need for community engagement mechanisms that the researchers expressed. Finally, the practitioner feedback corroborates Accessibility and Disciplinary Bridging (T6). While researchers surmised that PB primarily serves computational researchers, practitioners seemed to agree; for example, P9 observed "it's more for advanced users," and P14 noted that PB assumes "comfort with analytics, Python notebooks, and methodological thinking."

In terms of *complementing*, the practitioner feedback extends Template-Based Resources (T4) beyond what researchers articulated. In addition to templates, practitioners made other suggestions, such as "a small library/API, not just notebooks, for easier reuse" (P05) and "more diverse and downloadable sample personas" (P21). In addition, P11 suggested including "older datasets (belonging to the internet of a different time)". Finally, there was an interesting suggestion to incorporate evaluation and fairness considerations ("diversity checks to reduce bias in synthetic profiles", P23) that partially relate to the evaluation resources in PB but extend into data provenance and bias mitigation territory not directly addressed by the researchers.

In terms of *contradicting*, we did not find any feedback that would pose a contradiction with the researchers' feedback. Rather, practitioners and researchers seem to have slightly different emphases, but they largely gave similar feedback and suggestions to improve PB. Overall, the practitioner survey provides triangulating evidence for the researcher-derived themes, particularly around documentation, onboarding, and community building. It also adds practitioner-grounded specificity to the accessibility concerns and reveals additional considerations around dataset diversity and reusable infrastructure that could help PB provide more value for practitioners.

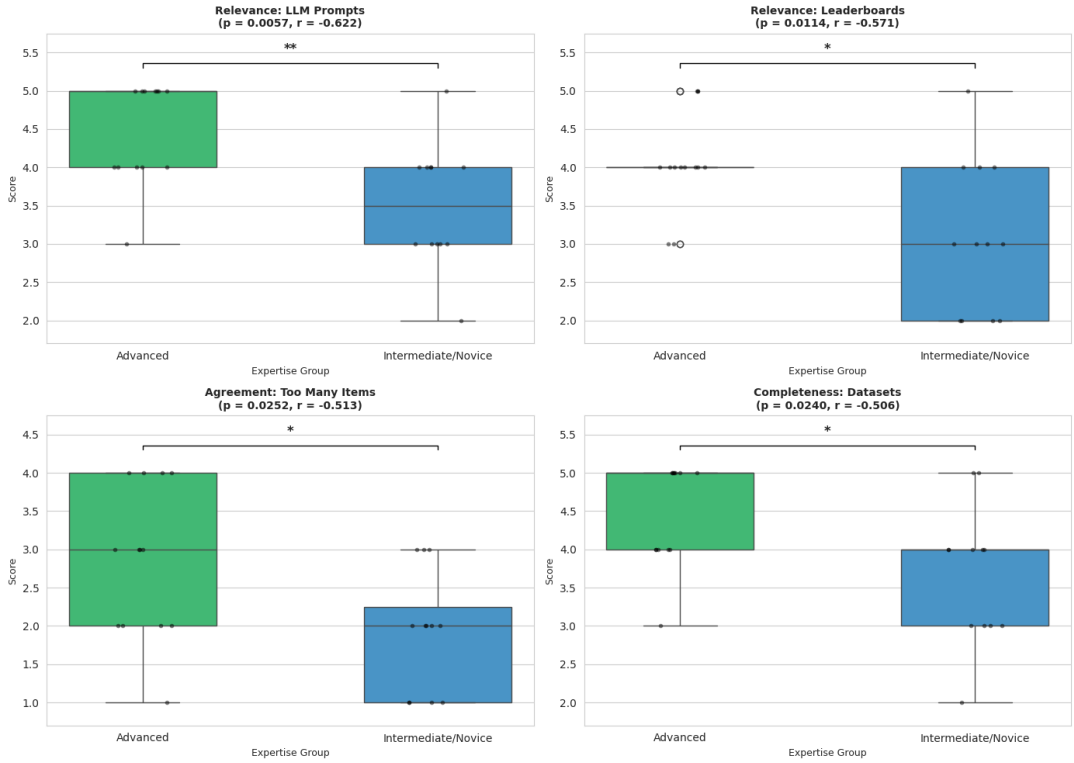


Figure 7: Significant effects of DDP experience.

5 Discussion

5.1 Research Implications

Our analysis suggests that innovation in DDP research has often benefited from a shared pool of resources. For example, researchers have used NMF [4, 41], developed as a general pattern recognition solution [45]. Similarly, other algorithms commonly used in persona creation, including k-means clustering, PCA, LDA, LSA, and so on (see review in [75]), have all been developed and made publicly available by researchers outside of the persona development domain and primarily have focused on problem spaces other than user representation or segmentation. Therefore, we can observe that persona research has *effectively appropriated* research and development in open science; yet, it has contributed little back and often applies methodologies that were not tailored for the specific purpose of persona development. Nevertheless, the development and sharing of resources would accelerate the cause of DDP research to make scientific progress as a field of study.

Returning to the critical reviewer quote we shared in the introduction, the work we have done for this current manuscript has led us to believe that the reviewer's concern is both valid and addressable. We agree that excessive 'benchmarkification' could lead to metric gaming rather than genuine progress if implemented carelessly. Persona research should avoid the pitfalls of focusing only on optimizing metrics, as personas are meant to enable understanding and empathy. Nevertheless, our findings suggest that the greater risk currently facing DDP research is not over-standardization but the absence of shared evaluation standards, which prevents cumulative knowledge building and makes it impossible to distinguish methodological progress from mere proliferation. Our formative evaluation with four persona researchers suggests that the current absence of computational resources actually hinders both novice and expert researchers, making it difficult to learn from previous work or determine whether new methods represent genuine improvements. On the other hand, the researchers articulated that thoughtfully designed computational resources could help democratize access to DDP methods while preserving—or even enhancing—the capacity for contextual, qualitative interpretation.

Therefore, the solution is not to completely avoid benchmarking, but to develop benchmarking practices that preserve qualitative richness while facilitating systematic comparison. For example, PB aims to support robust DDP practices while explicitly providing diverse metrics that help researchers avoid "single-metric optimization"—our leaderboard benchmark relies on four metrics focused on different aspects of fair and balanced DDP development, although we do not claim that PB would constitute a full benchmarking solution similar to GLUE in NLP.

How to get toward more complete benchmarking solutions, then? In their analysis of benchmarking in HCI, Hardy et al. [28] argue that effective benchmarks must offer realistic evaluations (1) grounded in real-world scenarios, (2) integrate relevant domain knowledge, (3) clearly communicate their scope and objectives, (4) assess varied tasks, (5) provide sufficient difficulty to prevent

rapid expiration, and (6) consider multiple performance dimensions instead of depending on singular metrics. An examination of PB indicates that the repository assesses varied tasks, including persona generation and evaluation, communicates its scope and objectives well enough for researchers and practitioners to understand, and includes multiple metrics to examine different aspects of persona generation. The applicability of the resources also covers real-world usage of DDPs, although the evaluators emphasized the importance of including more real-world datasets and mapping the resources more closely to specific scenarios and contexts. In terms of difficulty, while we introduce a leaderboard consisting of several persona evaluation metrics, testing the difficulty of achieving good performance using these metrics with a range of datasets remains a future work topic.

In addition to enabling the verification of scientific progress, benchmarking also *communicates* the state-of-the-art results in a field [28]. For example, given the lack of established benchmarks for DDP development, we cannot say which methods are better than others; and we cannot determine how far we are from reaching the best results. This challenge is associated, in part, with the "randomness" of the persona development. It has been observed, interestingly, that different algorithms create distinctly different personas from the same data [74]; similarly, different people create different personas from the same data [29]. Clearly, the ways in which the "black-box" thinking of people and machines is similar in their tendency to produce idiosyncratic results from the same baseline data. This poses a potential threat (or challenge) for benchmark building efforts in persona generation, or any form of user segmentation, for that matter. The crux of the issue is Chapman's classic notion that just as "pizza can be sliced in infinitely different ways" (paraphrasing [10]), so can data about people be sliced in very many different ways, and who is to say what "lens" applied is the correct one, if any? Although this may *seem* like an unsolvable problem, in theory, if we are able to define the requirements for "good personas" (or persona sets) and quantify these requirements, we can objectively compare personas obtained through different creation processes, human or machine-driven, and determine similarity between what was outputted and what we wanted.

5.2 Practical Implications

Based on our analysis, we propose three calls to action (CTAs) for the HCI research community.

CTA01: Contribute computational resources. Computational resources for DDPs could include, for example, (1) more formally defined tasks with clear objectives and boundary conditions, (2) a broader range of datasets corresponding to different practical use cases, (3) comprehensive metric libraries for benchmarking, and (4) standardized protocols such as reproducibility checklists for DDPs. For example, there are at least two opportunities to create additional datasets for DDPs: (a) common data structures and (b) synthetic data generation. The former refers to the fact that user data structures in dominant online platforms, such as Facebook, YouTube, Instagram, and TikTok are *known*, and these platforms often aggregate user data similarly [4], making it easier to create realistic *data templates* for benchmarking (see Figure 3). Using synthetic data generation, these templates could be populated with artificial (but realistically *distributed* [58]) data about users). When the purpose is to test different persona generation algorithms, for example, it does not matter if the data is real or synthetic[12], as long as it mimics realistic distributions and matches the available data structure in real systems.

Applying this logic, synthetic datasets [89] could be used for testing industry-compatible persona methods and systems even without primary access to specific accounts (that is, one needs to know the data structure but not the actual data in order to develop a general system for persona generation). For example, the APG system [40] contained a demo feature with synthetic data that would behave similarly to real YouTube Analytics data. Such resources should be centrally curated and published on commonly available platforms, including, for example, GitHub, Kaggle, Hugging Face, and the Open Science Foundation. The availability of resources can dramatically accelerate research and development by simplifying implementation and experimentation, as has been done in NLP. Furthermore, repositories should be kept current over time, as both persona methods and computational capabilities evolve, requiring continuous maintenance of repositories. A further step would be to package the recurring operations in the notebooks into a lightweight persona library with functions such as `generate()`, `enrich()`, and `evaluate()`, so that researchers can call shared implementations instead of copying code between notebooks. This requires coordination between DDP researchers, which we tackle in the subsequent call-to-action.

CTA02: Make community building a priority. Resource fortification should be approached as an iterative process of community building as opposed to a technical problem. However, "community" is not often discussed because DDP research tends to focus on documenting and reporting methods and findings. However, without a community that shares an understanding of tasks, datasets, metrics, and baselines, there cannot be a joint effort in progress. This implies that furthering resource contributions in DDPs is largely a *social mission* that requires social actions, such as events (e.g, workshops, challenges), publication opportunities (e.g., special issues, mini-tracks), and joint coordination (e.g., research projects, labs). Again, NLP is a great role model here. First, NLP conferences host a range of workshops focused on common tasks, such as SemEval[13], CLEF[14], and NTCIR[15].

CTA03: Disseminate knowledge across institutional and discipline barriers. Current persona development processes suffer from fragmented knowledge distribution, where theoretical insights from academic research remain disconnected from practical implementation tools, and domain-specific adaptations occur in isolation without cross-domain knowledge transfer [63]. To address

[12] Of course, it matters whether the data *behaves* realistically, which is why referred to the notion that the data distributions need to correspond to general population distributions.

[13] SemEval (Semantic Evaluation) is a series of international workshops that evaluate computational semantic analysis systems through standardized tasks, datasets, and evaluation metrics.

[14] CLEF (Conference and Labs of the Evaluation Forum, formerly Cross-Language Evaluation Forum) is a European initiative that promotes research in multilingual information access by organizing evaluation campaigns for information retrieval systems across different languages.

[15] NTCIR (NII Testbeds and Community for Information access Research) is a series of evaluation workshops designed to enhance research in information access technologies with a focus on East Asian languages and cross-language information retrieval

this fragmentation, knowledge dissemination is required. This is a communicative effort more than a technological one. It includes considering the varying *starting points* of people of different backgrounds. For example, using DDPs in companies involves several organizational roles, such as designers, researchers, developers, and managers that have different entry points and levels of methodological understanding. Ensuring the participation of multiple stakeholders needed for computational resource development requires a "buy in," which in turn requires communicating the need for these resources (and DDPs) in plain language. This could create more network effects [16] between scholars working on DDPs, it would accelerate development and experimental work, and it would perhaps attract more people to try out new techniques due to lower barriers of entry. All of these are desired positive outcomes that the field could use more of.

5.3 Future Research Directions

To continue evolving, DDP research needs foundational frameworks and principles that establish theoretical underpinnings, evaluation methods and tools that ensure quality and reliability, and practical implementation resources that facilitate adoption. Addressing these goals, we propose research directions (RDs) that we believe would benefit the DDP research community.

RD01: Draw inspiration from other computational fields. This could include, for example, analyzing successful NLP benchmarks, such as BLEU (37,952 citations in Google Scholar as of December 2025) and ROUGE (23,494 citations in Google Scholar as of December 2025) to investigate what made them successful in NLP and what could be learned from their success when building DDP benchmarks. Perhaps we should interview the creators of successful benchmarking resources in NLP to try and understand their interpretation of why the resources were successful, and suggest some broader "critical success factors" for benchmark building in computational sciences. In addition to drawing inspiration from the NLP and ML communities, computational research on DDPs should draw from the broader frameworks for responsible AI, including considerations of fairness, accountability, and transparency [44, 83]. In addition to concerns about rigor and reproducibility, DDP research should consider these various ethical aspects.

RD02: Investigate gaps in knowledge dissemination. There are at least two knowledge dissemination gaps that require in-depth investigation. First, the knowledge from DDP research to persona practice does not seem to transfer effectively. One can observe practitioners creating personas using poor methodologies and developing antipathy toward the persona technique as a consequence. Second, the knowledge between DDP research (HCI) and supporting fields (for example, NLP and ML) does not transfer effectively, as shown by the fact that common benchmarking practices in NLP are unfamiliar to HCI researchers, in general. For one, DDP development overlaps with benchmarking practices in specific computational tasks should be mapped—for example, clustering is well studied in computational benchmarking. What can be learned from those studies to make persona segmentation a benchmarkable task? Moreover, linking algorithmic diversity metrics to qualitative discussions of persona set completeness would help computational researchers understand theoretical motivations for technical choices. This future development work could help "cross-translate" each community's contributions to the other to avoid 're-trudging ground that earlier qualitative researchers have already done", as R03 pointed out. PB could potentially address this integration challenge by including curated references to qualitative persona literature, perhaps with annotations explaining how established principles relate to computational methods.

RD03: Define conceptual and theoretical boundaries for benchmarking in DDP research. Future research should also critically examine possible trade-offs between standardization and contextualization, quantitative and qualitative evaluation, and immediate versus longitudinal evaluation to ensure that benchmarking efforts advance rather than restricting the progress of DDP research. It is also worthwhile to note that benchmarking in other disciplines is not "perfect" either. The reader might have gotten the impression that everything is ideal in NLP, but in reality, benchmarking practices are under debate and periodically flux in NLP, too. For example, many benchmark datasets have become redundant or require updating in the wake of radical improvements in GenAI technologies. Therefore, benchmarking is not a "set and forget" activity but requires constant attention and corrective measures from the research community.

RD04: Develop experimental environments and cumulative evidence. Computational persona development processes benefit from interactive environments in which researchers can experiment with different approaches, observe system behaviors, and refine methodologies through iterative exploration. A key resource, and a potential advantage of HCI-specific interaction studies, are *platforms* that enable the flexible configuration of user experiments. For example, specific features could be turned on or off to examine the effect of specific features (e.g., explanations), or parameters for persona generation could be experimentally modified (e.g., generating more or fewer personas). An example of such platforms is illustrated by Jung et al. [39, 40] who provide, on one platform, (1) a functional system for persona users to generate and use personas based on their organization's primary data, (2) options for researchers to set up user studies according to specific interventions on the system, and (3) a persona analytics system that collects usage data from persona users and presents them to DDP researchers for observational or experimental studies.

Finally, we note that the current study does not demonstrate that PB supports real data-driven persona work, improves reproducibility, or enables robust benchmarking in practice. Our study was limited by the small expert sample, the brief practitioner exposure, and the limited representation of broader design practice. Rather than a mature benchmarking paper, our research should be understood as a curated, formative, and agenda-setting infrastructure contribution for DDP research. Therefore, more comprehensive evaluation studies of PB are needed, which we leave for future work.

6 Conclusion

Collectively, these findings indicate that PB addresses, although not complete fill, a genuine need for computational resources in the DDP research community by providing a well-organized consolidation of previously scattered resources. However, PB requires

improvement in documentation, reproducible end-to-end workflows, and active community engagement. These improvements would broaden its accessibility beyond technically proficient users and could help establish PB as a valuable tool for both academic and industry persona research.

We propose that persona research should develop in a way that does not exclude qualitative ideas and research traditions but carries these ideas when implementing new technologies in persona development. In turn, qualitative approaches could also benefit from adopting ideas from the computational side of personas. Overall, DDP methods need to have *empirical falsifiability*; they need to be able to be shown to be wrong. We need to be able to build increasingly accurate representations of reality by *verification* against empirical evidence, and the *outcomes* of our inquiry need to lead to knowledge and tools for the practitioner community, which plays a crucial role in DDP research as an applied research domain.

Declaration of Generative AI in the Writing Process

The authors used Claude and Gemini models to improve the content and its presentation style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for it.

Acknowledgments

We would like to thank Baki Kocaballi, Yoonseo Choi, and Jiangnan Xu for their valuable inputs that made this research possible.

References

[1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 4776–4790. doi:10.18653/v1/2023.emnlp-main.290

[2] Abeer Alessa and Hend Al-Khalifa. 2023. Towards Designing a ChatGPT Conversational Companion for Elderly People. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (2023) *(PETRA '23)*. Association for Computing Machinery, New York, NY, USA, 667–674. doi:10.1145/3594806.3596572

[3] Danial Amin, Joni Salminen, Farhan Ahmed, Sonja M.H. Tervola, Sankalp Sethi, and Bernard J. Jansen. 2026. Creating and Evaluating Personas Using Generative AI: A Scoping Review of 81 Articles. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3772318.3790608

[4] Jisun An, Haewoon Kwak, Soon-gyo Jung, Joni Salminen, M. Admad, and B. Jansen. 2018. Imaginary people representing real numbers: Generating personas from online social media data. *ACM Transactions on the Web (TWEB)* 12, 4 (2018), 1–26. ISBN: 1559-1131 Publisher: ACM New York, NY, USA. doi:10.1145/326598

[5] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642406

[6] Jonalan Brickey, Steven Walczak, and Tony Burgess. 2012. Comparing semi-automated clustering methods for persona development. *IEEE Transactions on Software Engineering* 38, 3 (2012), 537–546.

[7] Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis. *Proceedings of LREC 2022* (2022).

[8] Rudolf Carnap and Paul Arthur Schilpp. 1963. *The Philosophy of Rudolf Carnap*. Cambridge University Press, Cambridge.

[9] Chris Chapman and Elea McDonnell Feit. 2019. *R For Marketing Research and Analytics*. Springer.

[10] Chris Chapman, Kate Krontiris, and John Webb. 2015. *Profile CBC: Using Conjoint Analysis for Consumer Profiles*. Technical Report. Google Research.

[11] Chris Chapman, Edwin Love, Russell P. Milham, Paul ElRif, and James L. Alford. 2008. Quantitative Evaluation of Personas as Information. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 16 (Sept. 2008), 1107–1111. doi:10.1177/154193120805201602

[12] Chris Chapman and Russell P. Milham. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 5 (Oct. 2006), 634–636. doi:10.1177/154193120605000503

[13] Chaoran Chen, Leyang Li, Luke Cao, Yanfang Ye, Tianshi Li, Yaxing Yao, and Toby Jia-Jun Li. 2025. Why am I seeing this: Democratizing End User Auditing for Online Content Recommendations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 154, 23 pages. doi:10.1145/3746059.3747798

[14] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024. An Empathy-Based Sandbox Approach to Bridge the Privacy Gap among Attitudes, Goals, Knowledge, and Behaviors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–28. doi:10.1145/3613904.3642363

[15] Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent Trends in Personalized Dialogue Generation: A Review of Datasets, Methodologies, and Evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 13650–13665.

[16] Matthew T. Clements. 2004. Direct and indirect network effects: are they equivalent? *International Journal of Industrial Organization* 22, 5 (May 2004), 633–645. doi:10.1016/j.ijindorg.2004.01.003

[17] Alan Cooper. 1999. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity* (1 edition ed.). Sams - Pearson Education, Indianapolis, IN.

[18] S. De Paoli. 2023. Writing user personas with Large Language Models: Testing phase 6 of a Thematic Analysis of semi-structured interviews. *arXiv preprint arXiv:2305.18099* (2023).

[19] Stefano De Paoli. 2026. User personas, ideation and large language models: A post-hoc study. *International Journal of Human-Computer Studies* 208 (Jan. 2026), 103690. doi:10.1016/j.ijhcs.2025.103690

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[21] Elizabeth Ditton, Anne Swinbourne, and Trina Myers. 2022. Selecting a clustering algorithm: A semi-automated hyperparameter tuning framework for effective persona development. *Array* 14 (2022), 100186. Publisher: Elsevier.

[22] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one* 18, 3 (2023), e0279720. Publisher: Public Library of Science San Francisco, CA USA.

[23] Joy Goodman-Deane, Sam Waller, Dana Demin, Arantxa González-de Heredia, Mike Bradley, and John P. Clarkson. 2018. Evaluating Inclusivity using Quantitative Personas. In *In the Proceedings of Design Research Society Conference 2018* (2018). Limerick, Ireland. doi:10.21606/drs.2018.400

[24] Joy Ai-Leen Goodman-Deane, Mike Bradley, Sam Waller, and P. John Clarkson. 2021. Developing personas to help designers to understand digital exclusion. *Proceedings of the Design Society* 1 (Aug. 2021), 1203–1212. Publisher: Cambridge University Press. doi:10.1017/pds.2021.120

[25] Jonathan Grudin. 2006. Why personas work: The psychological evidence. *The Persona Lifecycle* (2006), 642–663.

[26] Jonathan Grudin and John Pruitt. 2002. Personas, participatory design and product development: An infrastructure for engagement. In *Proc. PDC* (2002), Vol. 2. 144–152.

[27] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3613904.3642472

[28] Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S. Bernstein, and Mykel John Kochenderfer. 2025. More than Marketing? On the Information Value of AI Benchmarks for Practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1032–1047. doi:10.1145/3708359.3712152

[29] Essi Häyhänen, Joni Salminen, and Bernard J. Jansen. 2025. Why Are Personas the Way They Are? Identifying Six Persona Creation Strategies. *Persona Studies* 11 (2025). doi:10.21153/psj2025vol11noart2002

[30] Pei-Fang Hsu, Yu-Han Lu, Shih-Chu Chen, and Patricia Pei-Yi Kuo. 2024. Creating and validating predictive personas for target marketing. *International Journal of Human-Computer Studies* 181 (Jan. 2024), 103147. doi:10.1016/j.ijhcs.2023.103147

[31] Zhengyu Hu, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Jianxun Lian, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025. Population-Aligned Persona Generation for LLM-based Social Simulation. arXiv:2509.10127 [cs]. doi:10.48550/arXiv.2509.10127

[32] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3544548.3580688

[33] Bernard Jansen, Soon-gyo Jung, and Joni Salminen. 2021. The Effect of Hyperparameter Selection on the Personification of Customer Population Data. *International Journal of Electrical and Computer Engineering Research* 1, 2 (2021). ISBN: 2791-6677.

[34] Bernard J. Jansen, Soon-Gyo Jung, Lene Nielsen, Kathleen W. Guan, and Joni Salminen. 2022. How to Create Personas: Three Persona Creation Methodologies with Implications for Practical Employment. *Pacific Asia Journal of the Association for Information Systems* 14, 3 (2022). doi:10.17705/1pais.14301

[35] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. 2020. From flat file to interface: Synthesis of personas and analytics for enhanced user understanding. *Proceedings of the Association for Information Science and Technology* 57, 1 (Oct. 2020). doi:10.1002/pra2.215

[36] Bernard J. Jansen, Joni Salminen, and Soon-Gyo Jung. 2020. Data-driven personas for enhanced user understanding: Combining empathy with rationality for better insights to analytics. *Data and Information Management* 4, 1 (2020), 1–17. ISBN: 2543-9251 Publisher: De Gruyter Poland. doi:10.2478/dim-2020-0005

[37] Soon-Gyo Jung, Jisun An, Haewoon Kwak, Moeed Ahmad, Lene Nielsen, and Bernard J. Jansen. 2017. Persona Generation from Aggregated Social Media Data. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017) *(CHI EA '17)*. ACM, New York, NY, USA, 1748–1755.

[38] Soon-Gyo Jung, Joni Salminen, Kholoud K. Aldous, and Bernard J. Jansen. 2025. PersonaCraft: Leveraging language models for data-driven persona development. *International Journal of Human-Computer Studies* 197 (2025), 103445. doi:10.1016/j.ijhcs.2025.103445

[39] Soon-Gyo Jung, Joni Salminen, and Bernard J. Jansen. 2021. Implementing Eye-Tracking for Persona Analytics. In *ACM Symposium on Eye Tracking Research and Applications*. 1–4.

[40] Soon-Gyo Jung, Joni Salminen, Haewoon Kwak, Jisun An, and Bernard J. Jansen. 2018. Automatic Persona Generation (APG) A Rationale and Demonstration. In *Proceedings of the 2018 conference on human information interaction & retrieval* (2018). 321–324. doi:10.1145/3176349.3176893

[41] Soon-Gyo Jung, Joni O. Salminen, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Automatically conceptualizing social media analytics data via personas. In *Twelfth International AAAI Conference on Web and Social Media* (2018). doi:10.1609/icwsm.v12i1.14992

[42] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Thi Thu Trang Xuan, Essi Häyhänen, Jinan Y. Azem, and Bernard J. Jansen. 2025. "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. In *Proceedings of the 30th International Conference on Intelligent User Interfaces* (2025). 1624–1638. doi:10.1145/3708359.3712160

[43] Ilkka Kaate, Joni Salminen, João M. Santos, Soon-Gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2024. "There Is something Rotten in Denmark": Investigating the Deepfake persona perceptions and their Implications for human-centered AI. *Computers in Human Behavior: Artificial Humans* 2, 1 (Jan. 2024), 100031. doi:10.1016/j.chbah.2023.100031

[44] Keith Kirkpatrick. 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Commun. ACM* 59, 10 (2016), 16–17.

[45] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (Oct. 1999), 788–791. doi:10.1038/44565

[46] Yunjo Lee, Grzegorz Migut, and Marcus Specht. 2023. What attention regulation behaviors tell us about learners in e-reading?: Adaptive data-driven persona development and application based on unsupervised learning. *IEEE Access* 11 (2023), 118890–118906.

[47] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 175–184. doi:10.18653/v1/2021.emnlp-demo.21

[48] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Position Paper Track*.

[49] C.-Y. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out* (2004). 74–81.

[50] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You Impress Me: Dialogue Generation via Mutual Persona Perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1417–1427. doi:10.18653/v1/2020.acl-main.131

[51] Nicola Marsden and Maren Haag. 2016. Stereotypes and politics: reflections on personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4017–4031.

[52] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012). ACM Press, Austin, Texas, USA, 1219. doi:10.1145/2207676.2208573

[53] Jennifer (Jen) McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008). ACM Press, Florence, Italy, 1521. doi:10.1145/1357054.1357292

[54] Eduard Metzker and Harald Reiterer. 2002. Use and Reuse of HCI Knowledge in the Software Development Lifecycle. In *Usability*, Judy Hammond, Tom Gross, and Janet Wesson (Eds.). Vol. 99. Springer US, Boston, MA, 39–55. Series Title: IFIP Advances in Information and Communication Technology. doi:10.1007/978-0-387-35610-5_3

[55] Tomasz Miaskiewicz, Tamara Sumner, and Kenneth A. Kozar. 2008. A latent semantic analysis methodology for the identification and creation of personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008). ACM, 1501–1510. http://dl.acm.org/citation.cfm?id=1357290

[56] T. Mijač, M. Jadrić, and M. Ćukušić. 2018. The potential and issues in data-driven development of web personas. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (2018). IEEE, 1237–1242. doi:10.23919/MIPRO.2018.8400224

[57] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

[58] Ha Nguyen, Victoria Nguyen, Saríah López-Fierro, Sara Ludovise, and Rossella Santagata. 2024. Simulating Climate Change Discussion with Large Language Models: Considerations for Science Communication at Scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*. Association for Computing Machinery, New York, NY, USA, 28–38. doi:10.1145/3657604.3662033

[59] Lan Huong Nguyen and Susan Holmes. 2019. Ten quick tips for effective dimensionality reduction. *PLoS computational biology* 15, 6 (2019), e1006907. Publisher: Public Library of Science San Francisco, CA USA.

[60] Lene Nielsen. 2018. Design personas - New ways, new contexts. *Persona Studies* 4, 2 (Nov. 2018), 1–4. doi:10.21153/psj2018vol4no2art799

[61] Lene Nielsen. 2004. *Engaging personas and narrative scenarios*. PhD Series, Vol. 17. Samfundslitteratur.

[62] Lene Nielsen. 2019. Going Global—International Personas. In *Personas - User Focused Design*, Lene Nielsen (Ed.). Springer, London, 123–133. doi:10.1007/978-1-4471-7427-1_7

[63] Lene Nielsen. 2019. *Personas - User Focused Design* (2nd ed. 2019 edition ed.). Springer, New York, NY.

[64] Lene Nielsen, Kira Storgaard Hansen, Jan Stage, and Jane Billestrup. 2015. A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *Int. J. Sociotechnology Knowl. Dev.* 7, 1 (Jan. 2015), 45–61. doi:10.4018/ijskd.2015010104

[65] James E. Nieters, Subbarao Ivaturi, and Iftikhar Ahmed. 2007. Making personas memorable. In *CHI '07 extended abstracts on Human factors in computing systems - CHI '07* (2007). ACM Press, San Jose, CA, USA, 1817. doi:10.1145/1240866.1240905

[66] Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[67] Shamik Panda. 2024. LLMs' ways of seeing User Personas. *arXiv preprint arXiv:2409.14858* (2024). doi:10.48550/arXiv.2409.14858

[68] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (2002). 311–318.

[69] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830. Publisher: JMLR. org.

[70] Karl Popper. 1957. Philosophy of science. *British Philosophy in the Mid-Century (ed. CA Mace). London: George Allen and Unwin* (1957).

[71] M. Prpa, G. Troiano, M. Wood, and Y. Coady. 2024. Challenges and Opportunities of LLM-Based Synthetic Personae and Data in HCI. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (2024) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–5. doi:10.1145/3613905.3636293

[72] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. doi:10.18653/v1/D16-1264

[73] Eric Rudolph, Natalie Engert, and Jens Albrecht. 2024. An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024)* (2024), Vol. 2. 108–117.

[74] Joni Salminen, Kamal Chhirang, Soon-Gyo Jung, Saravanan Thirumuruganathan, Kathleen W. Guan, and Bernard J. Jansen. 2022. Big Data, Small Personas: How Algorithms Shape the Demographic Representation of Data-Driven User Segments. *Big Data* (2022). doi:10.1089/big.2021.0177

[75] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2021. A Survey of 15 Years of Data-Driven Persona Development. *International Journal of Human–Computer Interaction* (2021), 1–24. ISBN: 1044-7318 Publisher: Taylor & Francis. doi:10.1080/10447318.2021.1908670

[76] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2022. Use Cases for Design Personas: A Systematic Review and New Frontiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2022). ACM, New Orleans LA USA, 1–21. doi:10.1145/3491102.3517589

[77] Joni Salminen, Soon-gyo Jung, Shammur Absar Chowdhury, Sercan Sengün, and Bernard J Jansen. 2020. Personas and Analytics: A Comparative User Study of Efficiency and Effectiveness for a User Identification Task. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2020). ACM, Honolulu, Hawaii, USA. doi:10.1145/3313831.3376770

[78] Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2022. Developing Persona Analytics Towards Persona Science. In *27th International Conference on Intelligent User Interfaces* (2022). ACM, 323–344. doi:10.1145/3490099.3511144

[79] Joni Salminen, Ilkka Kaate, Ahmed Mohamed Sayed Kamel, Soon-Gyo Jung, and Bernard J. Jansen. 2020. How Does Personification Impact Ad Performance and Empathy? An Experiment with Online Advertising. *International Journal of Human–Computer Interaction* 0, 0 (2020), 1–15. doi:10.1080/10447318.2020.1809246

[80] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J. Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642036

[81] Joni Salminen, Joao M. Santos, Haewoon Kwak, Jisun An, Soon-gyo Jung, and Bernard J. Jansen. 2020. Persona Perception Scale: Development and Exploratory Validation of an Instrument for Evaluating Individuals' Perceptions of Personas. *International Journal of Human-Computer Studies* 141 (2020), 102437. ISBN: 1071-5819 Publisher: Academic Press. doi:10.1016/j.ijhcs.2020.102437

[82] Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human-AI Workflows for Generating Personas. In *Designing Interactive Systems Conference* (2024). ACM, IT University of Copenhagen Denmark, 757–781. doi:10.1145/3643834.3660729

[83] Ying Shu, Jiehuang Zhang, and Han Yu. 2021. Fairness in Design: A Tool for Guidance in Ethical Artificial Intelligence Design. In *International Conference on Human-Computer Interaction* (2021). Springer, 500–510.

[84] Vandana Singh. 2019. Personas and Scenarios as a Methodology for Information Sciences. *Qualitative and Quantitative Methods in Libraries* 7, 1 (Feb. 2019), 123–134. Number: 1.

[85] Rashimi Sinha. 2003. Persona Development for Information-rich Domains. In *Extended abstracts of the 2003 Conference on Human Factors in Computing Systems (CHI 2003)*. Ft. Lauderdale, Florida, USA.

[86] Dimitris Spiliotopoulos, Dionisis Margaris, and Costas Vassilakis. 2020. Data-Assisted Persona Construction Using Social Media Data. *Big Data and Cognitive Computing* 4, 3 (2020), 21. doi:10.3390/bdcc4030021

[87] Igor Steinmacher, Marco Aurélio Graciotto Silva, and Marco Aurélio Gerosa. 2014. Barriers Faced by Newcomers to Open Source Projects: A Systematic Review. In *Open Source Software: Mobile Open Source Technologies*, Luis Corral, Alberto Sillitti, Giancarlo Succi, Jelena Vlasenko, and Anthony I. Wasserman (Eds.). Vol. 427. Springer Berlin Heidelberg, Berlin, Heidelberg, 153–163. Series Title: IFIP Advances in Information and Communication Technology. doi:10.1007/978-3-642-55128-4_21

[88] Seungjong Sun, Eungu Lee, Seo Yeon Baek, Seunghyun Hwang, Wonbyung Lee, Dongyan Nan, Bernard J. Jansen, and Jang Hyun Kim. 2024. Kiss up, Kick down: Exploring Behavioral Changes in Multi-modal Large Language Models with Assigned Visual Personas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 10888–10901. doi:10.18653/v1/2024.emnlp-main.609

[89] Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh R N, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025. PersonaBench: Evaluating AI Models on Understanding Personal Information through Accessing (Synthetic) Private User Data. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 878–893. doi:10.18653/v1/2025.findings-acl.49

[90] Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. 2022. Scientific machine learning benchmarks. *Nature Reviews Physics* 4, 6 (2022), 413–420.

[91] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems* (2019), Vol. 32.

[92] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 353–355. doi:10.18653/v1/W18-5446

[93] Chuanchi Wen, Paul Clough, Rachel Paton, and Rebecca Middleton. 2025. Leveraging large language models for thematic analysis: a case study in the charity sector. *AI & SOCIETY* (Aug. 2025). doi:10.1007/s00146-025-02487-4

[94] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3544548.3581388

[95] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2204–2213. doi:10.18653/v1/P18-1205

[96] Xishuo Zhang, Lin Liu, Yi Wang, Xiao Liu, Hailong Wang, Chetan Arora, Haichao Liu, Weijia Wang, and Thuong Hoang. 2024. Auto-Generated Personas: Enhancing User-centered Design Practices among University Students. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3613905.3651043

[97] Xishuo Zhang, Lin Liu, Yi Wang, Xiao Liu, Hailong Wang, Anqi Ren, and Chetan Arora. 2023. PersonaGen: A Tool for Generating Personas from User Feedback. In *2023 IEEE 31st International Requirements Engineering Conference (RE)*. 353–354. doi:10.1109/RE57278.2023.00048

[98] Haining Zhu, Honglu Wang, and John M. Carroll. 2019. Creating Persona Skeletons from Imbalanced Datasets–A Case Study using U.S. Older Adults' Health Data. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (2019). 61–70. doi:10.1145/3322276.3322285

[99] Luyao Zhu, Wei Li, Rui Mao, and Erik Cambria. 2024. HIPPL: Hierarchical Intent-Inferring Pointer Network With Pseudo Labeling for Consistent Persona-Driven Dialogue Generation [Research Frontier]. *IEEE Computational Intelligence Magazine* 19, 4 (2024), 63–78. Publisher: IEEE.