Person Distillation: Foundations and Principles for Personalized AI

MINXING ZHANG*, Duke University, USA
YI YANG*, Duke University, USA
JIAN PEI, Duke University, USA

Large language models are rapidly evolving from general-purpose assistants into personalized AI systems that collaborate with individuals over long periods of time. Emerging applications increasingly seek not only to answer questions, but also to reason, communicate, and act in ways that faithfully reflect a particular person. Recent systems have begun to distill personal knowledge, workplace experience, communication history, and behavioral traces into AI assistants, skill libraries, memory systems, and personalized agents. Despite this rapid progress, the field remains fragmented. Existing work differs substantially in the evidence it uses, the aspects of a person it preserves, the representations it constructs, the methods it employs, and the way it evaluates fidelity, leaving no common conceptual foundation for person distillation.

This paper presents the first comprehensive framework for *person distillation*: the problem of transforming heterogeneous traces of a particular real person into faithful, evidence-grounded, and revisable computational representations. We organize the field as an end-to-end lifecycle consisting of four tightly coupled components: *source evidence*, *distilled person representations*, *distillation methods*, and *evaluation*. For each component, we develop a taxonomy that unifies existing research and identifies open design choices. We further organize the methodological landscape into six complementary families, ranging from prompt-based trace-to-profile distillation to memory-based, parametric and adapter-based, preference- and reward-based, behavioral and trajectory-based, and hybrid and revisable distillation. Finally, we propose a comprehensive evaluation framework covering person-fidelity objectives, benchmark construction protocols, quantitative metrics, and diagnostic analyses for measuring evidence grounding, person specificity, boundary awareness, and evaluation reliability.

Beyond surveying existing work, this paper establishes a common vocabulary, a conceptual framework, and a research agenda for person distillation. By connecting advances in role-playing LLMs, personalized generation, agent memory, skill learning, and preference modeling, we identify the fundamental challenges that distinguish modeling a *particular real person* from modeling generic users, roles, or personas. As AI systems become increasingly personalized, collaborative, and long-lived, we believe person distillation will become a foundational capability for the next generation of AI assistants, digital coworkers, and human-centered intelligent systems.

1 Introduction

Large language models (LLMs) are rapidly evolving from general-purpose assistants into long-term collaborators that work with people over extended periods of time. Beyond answering questions, future AI systems will help individuals write papers, review code, make decisions, manage projects, and carry out complex workflows across months or even years [59]. Achieving this vision requires more than stronger reasoning or better personalization. AI systems must understand particular people: what they know, how they reason, how they communicate, and how they make decisions. We call this capability *person distillation* and believe it will become a foundational component of the next generation of personalized AI.

Current personalization captures only part of this goal. Existing systems typically remember user information, adapt communication style, or learn task-specific preferences [98, 120]. They improve user experience, but they do not faithfully represent an individual. A personalized assistant may learn that an engineer prefers concise explanations,

*Minxing Zhang and Yi Yang contributed equally to this work.

Authors' Contact Information: Minxing Zhang, minxing.zhang@duke.edu, Duke University, Durham, North Carolina, USA; Yi Yang, owen.yang@duke.edu, Duke University, Durham, North Carolina, USA; Jian Pei, j.pei@duke.edu, Duke University, Durham, North Carolina, USA.

yet still fail to capture how that engineer evaluates deployment risk, balances competing objectives, or decides when evidence is sufficient. Faithfully representing a person therefore requires a substantially richer notion of personalization.

We define *person distillation* as the process of transforming heterogeneous traces of a particular real person into evidence-grounded, person-specific, and revisable computational representations for specified uses. Unlike conventional *knowledge distillation* [22, 34], which transfers capability from a teacher model to a student model, person distillation begins with evidence about a real individual and seeks to preserve person-specific knowledge, judgment, communication, behavior, and other characteristics. The resulting representation may take many forms, including document profiles, memory systems, parametric models, preference or reward models, and hybrid combinations thereof.

Faithfully representing a person is fundamentally different from imitating one. A system may reproduce a person's writing style while failing to preserve their expertise, make technically correct decisions that differ from the person's judgment, or invent unsupported preferences, experiences, or decision rules. Person distillation therefore asks three fundamental questions: *What* information about a person is supported by the available evidence? *What* makes that information specific to the individual rather than the role? *When* should the system acknowledge that the evidence is insufficient?

These questions arise naturally as AI systems become long-term collaborators. Consider a senior software engineer who has spent years developing a large-scale payment infrastructure. Existing documentation records which services the engineer maintained and which design documents they authored, but rarely captures how they evaluated deployment risk, what evidence they required before approving a launch, or how they communicated technical tradeoffs to different teammates. Supporting future collaborators or AI assistants therefore requires more than a generic "senior backend engineer" persona. It requires a computational representation that faithfully preserves person-specific expertise, judgment, and behavior while remaining grounded in observable evidence. Figure 1 illustrates this vision, while Figure 2 previews the end-to-end lifecycle that this paper develops.

Although person distillation builds upon several active research directions, it is not reducible to any one of them. Existing work has made substantial progress in role-playing, personalization, and skill learning, yet these directions optimize different objectives. Role-playing models aim to emulate identities or personas [56, 98, 120, 123, 132]; personalized generation adapts models to individual users and tasks [94, 136]; and agent skill learning distills reusable capabilities from trajectories and experience [41, 77, 118]. Person distillation addresses a different problem: constructing faithful computational representations of a particular real person from heterogeneous evidence.

The distinction is fundamental. Existing research primarily models one of three objects: *how a person appears*, *what a user prefers*, or *how a task is accomplished*. Person distillation instead models *a particular real person*, including their knowledge, judgment, communication, behavior, and relationships, while remaining grounded in available evidence. This broader objective introduces new requirements, including evidence grounding, person specificity, temporal evolution, revisability, and explicit recognition of the limits of available evidence.

Recent research systems and open-source projects demonstrate that this vision is increasingly practical [1, 3, 23, 109, 116, 126, 135]. They distill workplace experience into colleague skills, transform communication histories into personal assistants, construct expert agents from professional artifacts, and build developer portraits from interactions with AI coding assistants. Collectively, these systems show that heterogeneous traces of real people can be distilled into useful computational representations.

However, the field remains fragmented. Existing systems differ substantially in the evidence they use, the aspects of a person they preserve, the representations they construct, the methods they employ, and the way they evaluate fidelity. Most are designed for specific applications, rely on prompt-based profile generation, and lack common datasets,

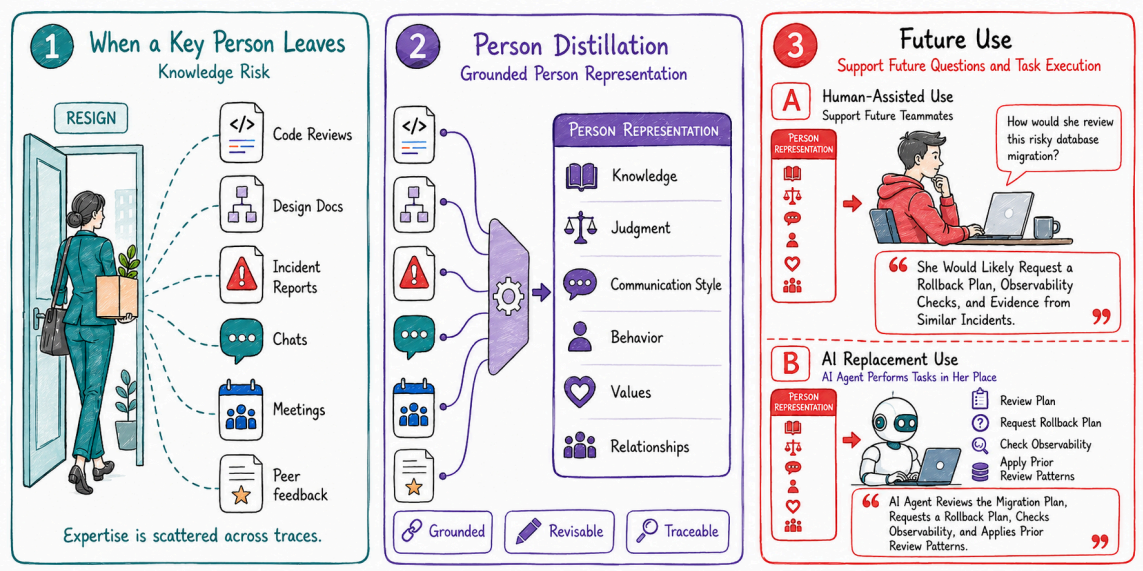

Fig. 1. Motivation for person distillation. When a key person leaves an organization, person-specific domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context remain distributed across heterogeneous traces. Person distillation transforms these traces into evidence-grounded and revisable representations, such as profiles, memory systems, adapters, reward models, or hybrid representations. These representations can support future collaborators or guide an AI agent in bounded, person-specific tasks.

evaluation protocols, or benchmark tasks. Consequently, existing work demonstrates that person-related representations can be constructed, but it remains unclear what has actually been distilled, how faithfully it has been preserved, or how different approaches should be compared.

This fragmentation reflects a deeper challenge: person distillation lacks a common conceptual foundation. Designing or evaluating a distilled representation requires jointly answering four tightly coupled questions: What evidence is available? What aspects of the person should be preserved? How should they be represented? How should fidelity be evaluated? These questions define an end-to-end lifecycle rather than independent design decisions.

We therefore organize person distillation as a unified lifecycle consisting of four connected components: *source evidence*, *distilled person representations*, *distillation methods*, and *evaluation* (Figure 2). This lifecycle provides a common conceptual framework for organizing existing systems, clarifying their relationships, and identifying unexplored research opportunities. Rather than viewing document profiles, memories, parametric models, and reward models as competing alternatives, we treat them as complementary representations within a single person-distillation pipeline.

Building on this framework, this paper establishes the foundations of person distillation as an emerging research area. We develop a taxonomy for each stage of the lifecycle, organize existing methodologies into six complementary families, and present a comprehensive evaluation framework for measuring person fidelity. Together, these components provide a common vocabulary and a reference framework for designing, comparing, and evaluating future person-distillation systems.

Our main contributions are summarized as follows.

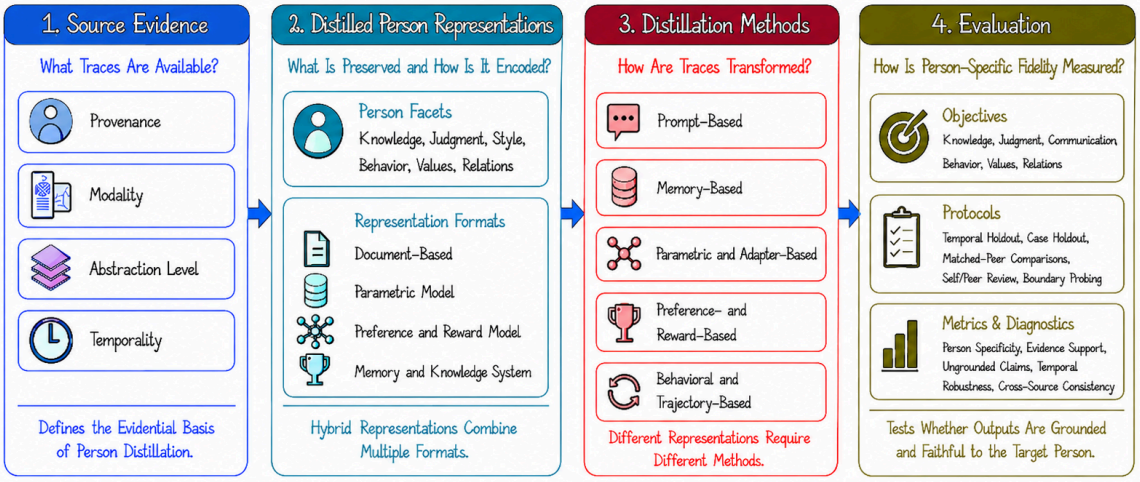

Fig. 2. A general taxonomy of person distillation. The taxonomy organizes person distillation as a coupled lifecycle from source evidence to distilled person representations, distillation methods, and evaluation.

- We define *person distillation* as the process of transforming heterogeneous traces of a particular real person into evidence-grounded, person-specific, and revisable computational representations, and distinguish it from knowledge distillation, role-playing, personalization, memory-based assistants, and agent skill distillation.
- We establish the first end-to-end lifecycle for person distillation, organized around four connected components: source evidence, distilled person representations, distillation methods, and evaluation, with a taxonomy for each stage.
- We organize existing methodologies into six complementary families and analyze how they preserve different facets of a person, providing a unified view of the emerging methodological landscape.
- We present a comprehensive evaluation framework covering objectives, protocols, metrics, and diagnostics for measuring person fidelity, evidence grounding, person specificity, boundary awareness, and evaluation reliability.

We believe person distillation will become a foundational capability for personalized AI. As AI systems evolve from generic assistants into long-term collaborators, they must move beyond remembering isolated preferences or imitating conversational style toward faithfully representing particular people. We hope this paper provides a common conceptual foundation for this emerging research area.

The remainder of this paper is organized as follows. Section 2 presents the proposed lifecycle and taxonomy. Section 3 examines personal traces as source evidence. Section 4 studies person representations. Section 5 reviews distillation methods. Section 6 develops the evaluation framework. Finally, we discuss open challenges and future research directions in Section 7.

2 A Taxonomy of Person Distillation

Person distillation spans the entire pipeline from personal traces to faithful computational representations, yet existing work has largely studied its individual components in isolation [23, 77, 94, 98, 120, 135]. A unified taxonomy is therefore

needed to organize the design space, clarify how different research directions relate to one another, and position both existing systems and future work within a common framework.

We organize person distillation as a *hierarchical taxonomy*. At the top level, person distillation is viewed as an end-to-end lifecycle consisting of four tightly connected components: *source evidence*, *distilled person representations*, *distillation methods*, and *evaluation*. At the second level, each component is further developed into its own taxonomy, capturing the major design dimensions and research challenges associated with that stage. Figure 2 illustrates this overall organization.

2.1 Overview of the Taxonomy

Figure 2 presents the top-level taxonomy of person distillation as an end-to-end lifecycle. The lifecycle begins with *source evidence*: the personal traces from which information about a person can be inferred. These traces are distilled into *person representations* that preserve different aspects of the individual, such as knowledge, judgment, communication style, behavior, values, and relationships. *Distillation methods* transform heterogeneous traces into these representations, while *evaluation* measures whether the resulting representation faithfully captures the target person. These four components jointly define the overall design space of person distillation.

These components are tightly coupled rather than independent. The available evidence determines which claims about a person can be supported. The aspects of the person that should be preserved determine the appropriate representation. The chosen representation shapes the corresponding distillation methodology, and the evaluation protocol must ultimately assess whether the representation faithfully captures the intended person. Consequently, person distillation cannot be understood by studying any single component in isolation; the entire lifecycle must be considered as a coherent system.

This taxonomy is guided by several principles. Person distillation should be *evidence-grounded*, so that person-specific claims remain supported by observable traces. It should be *facet-aware*, recognizing that knowledge, judgment, communication, behavior, values, and relationships are distinct aspects of a person. It should be *representation-aware*, allowing different facets to be encoded using different computational forms. It should be *temporally aware*, accounting for how people evolve over time. Finally, it should remain *person-specific*, preserving what distinguishes one individual from others with similar roles or expertise.

The remainder of this section develops a dedicated taxonomy for each component of this lifecycle.

2.2 Source Evidence

The first component of person distillation is *source evidence*: the personal traces from which a person's knowledge, behavior, preferences, relationships, and other characteristics can be inferred [54, 101]. Since different traces support different kinds of claims, source evidence defines the boundary of what can be responsibly distilled [21, 114]. For example, a diary entry, a meeting recording, and a peer evaluation may all describe the same individual, but each provides different evidence about that person. We organize source evidence along four complementary dimensions: *provenance*, *modality*, *abstraction level*, and *temporality*. Together, these dimensions characterize what evidence is available and what person-specific claims it can support.

Provenance describes who generated a trace. *First-person* traces are produced by or directly attributable to the target person, such as messages, emails, code, design notes, voice memos, calendar events, or activity logs [135]. *Third-party* traces are produced by others who observe, evaluate, interact with, or discuss the target person, such as peer reviews, recommendation letters, customer feedback, media reports, or conversations about the person [21, 114].

These two sources provide complementary evidence. First-person traces reveal what the person actually said, wrote, chose, or did, although they may be influenced by context or self-presentation. Third-party traces capture how the person is perceived by others, but they may also reflect observer bias or misunderstanding. Provenance therefore determines whether a claim is directly observed, externally perceived, or requires evidence from multiple perspectives.

Modality describes how a trace is expressed and organized. Personal traces may take the form of text, speech, video, images, code, behavioral logs, task trajectories, relational graphs, or combinations of these modalities [23, 135]. They also range from unstructured messages and recordings to semi-structured emails and tickets, and fully structured logs, timelines, and networks [77].

Different modalities expose different aspects of a person. Text often reveals expertise and communication style [83], audio and video capture conversational dynamics, and relational graphs expose collaboration, influence, and mentorship [4]. Consequently, the available modalities determine which aspects of a person are observable and which remain hidden.

Abstraction level describes how much interpretation has already been applied to a trace. *Raw traces*, such as original messages, recordings, documents, logs, or actions, preserve detailed behavioral evidence but are often noisy and difficult to analyze directly [101]. *Interpreted traces*, such as biographies, performance reviews, summaries, recommendation letters, or personality assessments, provide compact descriptions but already incorporate human or algorithmic interpretation [114].

This distinction creates a tradeoff between fidelity and compression. Raw traces preserve subtle behavioral patterns that may never be stated explicitly, whereas interpreted traces are easier to use but inevitably lose detail and may reflect the assumptions of their authors [86]. Abstraction level therefore determines how directly a claim is grounded in observable behavior.

Temporality describes when a trace was produced and how its temporal structure is represented. Some traces are *snapshots*, such as resumes, interviews, public profiles, or annual reviews. Others are *longitudinal*, including communication archives, project histories, collaboration records, or social media timelines.

Temporality is essential because people evolve over time. Knowledge accumulates, roles change, preferences drift, relationships develop, and behaviors adapt to new contexts [89]. Ignoring time may merge outdated and current evidence into a single static portrait, whereas relying only on recent traces may overlook enduring expertise or values. Temporality therefore determines which version of a person is being represented and which characteristics are stable versus evolving.

Taken together, these four dimensions provide a structured view of source evidence. They clarify what person-specific information can be inferred from available traces, what claims are supported by evidence, and what representations, methods, and evaluation protocols are appropriate for the later stages of person distillation.

2.3 Distilled Person Representations

The second component of person distillation concerns the *representation* produced by the distillation process. Once person-specific information has been extracted from personal traces, a system must determine both *what* aspects of the person should be preserved and *how* they should be represented. Correspondingly, we organize the representation design space along two complementary dimensions: *person facets*, which describe the content being preserved, and *representation formats*, which describe how that content is computationally encoded. These two dimensions are orthogonal: the same facet can be represented in different formats, and the same representation format can support multiple facets.

Person facets describe what aspects of a person the representation aims to preserve. These facets include domain knowledge and expertise [28], judgment and decision criteria [44], communication style [83], behavioral patterns [73], values and personality [67], and relational context [4]. They capture different dimensions of a person and should not be collapsed into a single generic persona [56, 98, 120]. Two individuals may possess similar expertise but make different decisions, communicate differently, or maintain different relationships with others. Organizing representations by facets therefore clarifies both what a system preserves and what remains outside its scope.

Representation formats describe how person-specific information is encoded. A distilled person may be represented as an explicit document, such as a profile, skill file, system prompt, or behavioral specification [120, 135]. It may also be represented as a parametric model or adapter that learns recurring patterns [37, 98], a preference or reward model that captures evaluative judgment [10, 20], or a memory system that stores retrievable experiences together with their supporting evidence [81].

Each format offers different strengths and limitations. Documents are transparent and easy to revise, but may miss tacit knowledge [78]. Parametric models can learn implicit regularities, but are difficult to inspect [26]. Reward models naturally capture judgment, but only evaluate candidate alternatives [10]. Memory systems preserve evidence and support traceability, but require effective retrieval and reasoning [81]. Distinguishing representation formats therefore clarifies the computational object that a distillation method is expected to produce.

Person facets and representation formats do not have a one-to-one correspondence. A single facet may be represented in multiple formats, while one representation format may support multiple facets. Consequently, faithful person distillation is unlikely to rely on a single profile, model, reward function, or memory store [10, 37, 81]. Instead, it will likely require a **hybrid representation** that combines multiple formats, allowing each to preserve the aspects of a person for which it is best suited.

Hybrid representations are also motivated by the multimodal nature of personal traces. Different modalities, such as text, code, speech, video, behavioral logs, and relational graphs, often reveal complementary aspects of a person [23, 135], and no single representation format is equally effective at preserving all of them. A hybrid representation therefore provides a natural mechanism for integrating heterogeneous evidence while maintaining the strengths of different computational representations.

2.4 Distillation Methods

The third component of person distillation concerns *how* distilled person representations are constructed from personal traces. Given heterogeneous traces and a target representation, a distillation method must determine what information to extract, how it should be abstracted, and how it should be represented.

We organize the methodology design space according to the primary *computational mechanism* used to preserve person-specific information. Different methods encode different aspects of a person in different ways. Some summarize traces into explicit textual descriptions [120, 135]; others organize them as retrievable memories [81]; others learn implicit behavioral patterns in model parameters [37, 98]; still others represent evaluative judgment [10, 20] or procedural behavior [77]. These different mechanisms make different assumptions about evidence, supervision, generalization, interpretability, and revision, giving rise to six complementary families of distillation methods.

Prompt-Based Trace-to-Profile Distillation methods use large language models to summarize personal traces into a *single* explicit textual representation, such as a persona description, skill file, behavioral summary, or structured profile [120, 135]. This approach performs global compression by distilling diverse evidence into one document that is

simple, transparent, and easy to integrate into existing LLM agents. However, such compression may discard subtle behavioral patterns, implicit decision criteria, contextual information, and links to supporting evidence [78, 86].

Memory-Based Distillation methods address this limitation by replacing a single global profile with a collection of *distilled memory units*. Instead of compressing all traces into one document, they organize episodes, facts, procedures, relationships, timestamps, and supporting evidence into a queryable memory system [55, 81]. Each memory captures a small, coherent piece of person-specific knowledge while preserving its connection to the underlying evidence, enabling grounded and traceable reasoning. Their main challenge lies in deciding what memories to construct, how to organize and retrieve them, and how to generalize from historical experiences to new situations.

Parametric and Adapter-Based Distillation methods encode person-specific patterns directly into model parameters, adapters, embeddings, or other learned components [37, 94, 98]. They are well suited for capturing recurring reasoning patterns and behavioral tendencies that are difficult to express explicitly [78]. Their main limitation is reduced interpretability, making it difficult to understand what has been learned, detect memorization of sensitive information, or verify that the learned behavior truly reflects the target person [15, 26].

Preference- and Reward-Based Distillation methods represent a person through their evaluative behavior by learning how they compare or rank alternative decisions [20, 80]. This representation is particularly effective for modeling judgment and decision criteria [10]. However, evaluating alternatives is not equivalent to generating them, and reducing complex reasoning to scalar preference scores may obscure the rationale behind a person's decisions.

Behavioral and Trajectory-Based Distillation methods learn from sequences of actions rather than isolated artifacts [77, 111]. They model how a person gathers information, makes decisions, revises plans, and responds to feedback across time. Such methods are important for capturing procedural behavior, but they must address the challenge that real-world human trajectories are often fragmented, incomplete, and distributed across multiple contexts and platforms.

These methodological families are complementary rather than mutually exclusive. A practical person-distillation system may combine document-based profiles for explicit knowledge, memory systems for evidence, parametric models for implicit behavioral patterns, and reward models for evaluative judgment [10, 37, 81, 135]. An important research challenge is therefore to develop **Hybrid and Revisable Person Distillation** methods that coordinate multiple representations, integrate heterogeneous evidence, resolve inconsistencies, and continually update the distilled representation as new evidence becomes available [70, 84].

2.5 Evaluation

The fourth component of person distillation concerns *evaluation*. Once a distilled person representation has been constructed, the system must determine whether it faithfully captures the target person. Unlike conventional AI evaluation, person distillation cannot be assessed solely by task success, response quality, or user satisfaction [64, 94]. A system may complete a task successfully while failing to reflect the person's knowledge, judgment, communication style, or behavior [120]. It may also generate convincing but unsupported person-specific claims [40]. We therefore organize the evaluation design space around three complementary dimensions: *evaluation objectives*, *protocol design*, and *metrics and diagnostics*.

Evaluation objectives define what aspect of person fidelity is being measured. Because a distilled person representation may encode multiple facets, evaluation should distinguish knowledge fidelity, judgment fidelity, communication fidelity, behavioral fidelity, values and personality fidelity, and relational fidelity. These objectives capture different aspects of a person and should be evaluated separately rather than collapsed into a single score.

Protocol design specifies how evaluation cases are constructed. Person distillation requires protocols that assess both generalization and person specificity. For example, temporal hold-out evaluates whether historical traces predict future behavior [39], case hold-out measures generalization to unseen situations [50], matched-peer comparison tests whether the system distinguishes the target person from similar individuals [102], self and peer review assesses perceived faithfulness [114], and boundary probing examines whether the system appropriately abstains when evidence is insufficient [31, 40].

Metrics and diagnostics determine how evaluation results are measured and interpreted. Depending on the target facet, evaluation may measure factual accuracy over known experiences, agreement with historical decisions, similarity of communication patterns, alignment with observed behavioral trajectories, or consistency with peer assessments [102]. Person distillation also requires diagnostics beyond conventional model evaluation, including person specificity relative to matched baselines, evidence support for generated claims, unsupported claim rate, boundary calibration, temporal robustness, and cross-source consistency [31, 40].

Section 6 develops these three dimensions into a comprehensive evaluation framework, including objectives, benchmark construction protocols, metrics, and diagnostic analyses for assessing person-specific fidelity.

3 Source Evidence: A Taxonomy of Personal Traces

Person distillation begins with source evidence. Before asking how to extract, represent, or evaluate a person's knowledge and behavior, we must first ask what evidence about that person is available. The coverage, granularity, provenance, modality, abstraction level, and temporal structure of this evidence fundamentally determine what can be distilled. A system trained only on chat messages may imitate conversational style but is unlikely to recover technical judgment or decision-making patterns. A system trained only on task trajectories may capture reusable skills but miss the person-specific reasoning behind those skills.

We define **personal traces** as observable evidence from which a person's knowledge, skills, preferences, personality, relationships, and behavioral patterns can be inferred. Personal traces include *first-person artifacts* such as messages, emails, code commits, authored documents, meeting recordings, social media posts, and activity logs, as well as *third-party artifacts* such as peer evaluations, performance reviews, media reports, and conversations about the target person. Personal traces are therefore not a single data type but a heterogeneous collection of signals accumulated across personal, professional, and social contexts.

To organize this diversity, we analyze personal traces along four dimensions: *provenance* (Section 3.1), which concerns who generated the trace; *modality* (Section 3.2), which concerns how the trace is expressed; *abstraction level* (Section 3.3), which concerns how much interpretation has already been applied; and *temporality* (Section 3.4), which concerns when the trace was generated and whether its temporal structure is preserved. These dimensions are complementary. A trace can simultaneously be first-person, textual, raw, and timestamped; another can be third-party, visual, highly interpreted, and retrospective. Figure 3 summarizes this taxonomy and provides illustrative trace examples.

This taxonomy also reveals an important limitation of existing work. Role-playing and persona-based LLMs [56, 98, 120, 123, 132] primarily rely on curated dialogues, scripts, and persona descriptions, emphasizing conversational style and character consistency. Agent skill distillation frameworks [41, 77, 118] primarily rely on task execution trajectories and feedback, emphasizing task success rather than person-specific reasoning. Existing person distillation systems [23, 109, 116, 135] have begun to incorporate richer personal traces, but they still cover only a small portion of the broader source-data landscape.

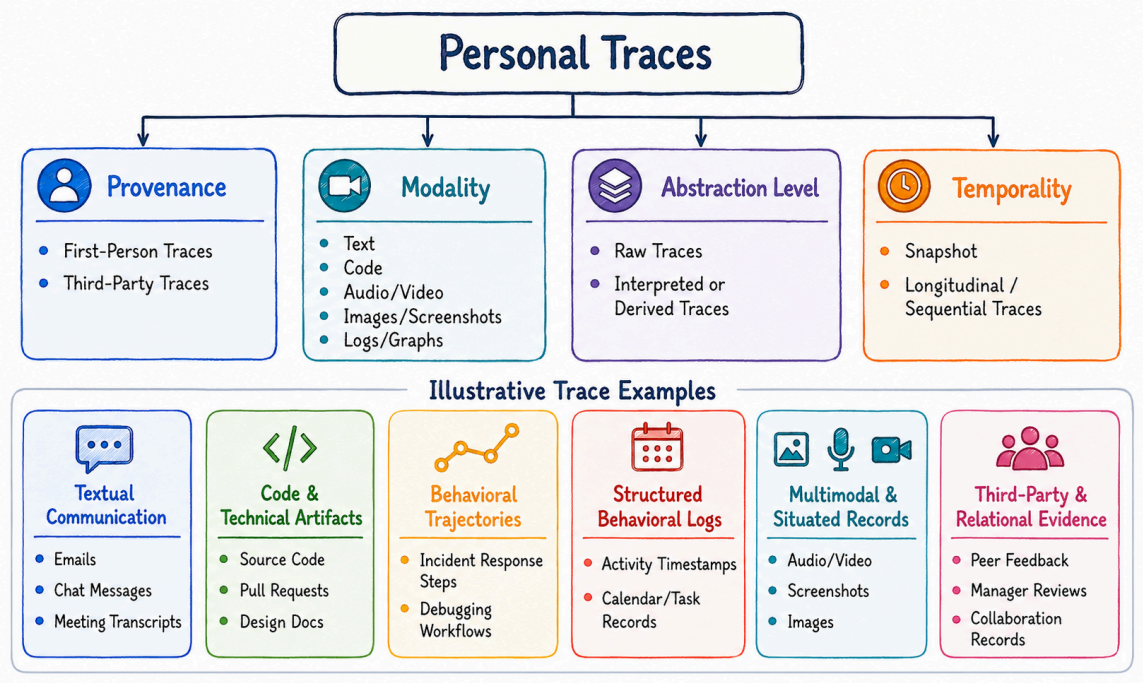

Fig. 3. Taxonomy of personal traces. The top panel shows the four organizing dimensions: provenance, modality, abstraction level, and temporality. The bottom panel gives illustrative trace-type examples.

The remainder of this section develops the taxonomy and highlights research opportunities associated with each dimension.

3.1 Trace Provenance: Who Generated the Evidence?

Provenance concerns who generated a trace: the target person or others who observed, interacted with, evaluated, or discussed that person. Provenance matters because different sources reveal different aspects of an individual. First-person traces provide direct evidence of what a person said, wrote, built, chose, and did. Third-party traces provide evidence of how that person was perceived, evaluated, and experienced by others.

3.1.1 First-Person Traces. **First-person traces** are artifacts produced by or directly attributable to the target person. Examples include chat messages, emails, code commits, authored documents, social media posts, meeting recordings, voice memos, screenshots, and activity logs. Existing person distillation systems rely heavily on such traces. For example, Colleague-Skill [135] ingests a colleague's Feishu chat messages, DingTalk documents, PDFs, images, and screenshots. Similarly, Ex-Skill [109] collects WeChat and iMessage conversation logs, screenshots, and related personal artifacts.

The primary advantage of first-person traces is direct attribution. The target person actually wrote the message, committed the code, authored the document, or made the decision. As a result, these traces provide high-fidelity evidence of behavior with minimal intermediary interpretation. They can reveal communication style, technical preferences, decision patterns, routines, expertise, and values.

Evidence from computational social science supports this view. Kosinski et al. [54] show that Facebook Likes can predict personal attributes such as personality traits, political views, and sexual orientation. Stachl et al. [101] demonstrate that smartphone behavioral data, including app usage, communication logs, mobility patterns, and screen activity, can predict Big Five personality traits [68]. In software engineering, Caliskan-Islam et al. [13] show that programmers can be de-anonymized from executable binaries with up to 96% accuracy, even after compilation removes variable names and alters program structure. These results suggest that first-person traces encode not only observable behavior but also deeper person-specific patterns of reasoning, choice, and judgment.

First-person traces are also attractive because they are naturally generated in modern digital environments. Messages, emails, documents, commits, calendar events, and activity logs accumulate as a byproduct of everyday work and life. However, abundance does not imply informativeness. Many traces are routine, repetitive, context-dependent, or only weakly related to a person's distinctive characteristics. A message such as "sounds good" is clearly attributable to the individual yet reveals little about their expertise or judgment.

Moreover, first-person traces are often shaped by self-presentation. People curate public posts, write formal emails differently from private notes, and may alter their behavior when they know they are being observed. Consequently, first-person traces are not neutral observations of a person. They are direct evidence of behavior, but they may overrepresent what the individual chose to record, communicate, or display.

Finally, first-person traces raise significant privacy concerns. Because they originate from the target person, they may contain sensitive information about preferences, relationships, health, politics, location, work history, private conversations, or confidential projects. The same property that makes these traces valuable for person distillation, namely, their ability to encode rich personal information, also makes them risky to collect, store, model, and expose. Person distillation systems therefore require mechanisms for provenance tracking, consent management, access control, temporal scoping, selective forgetting, and the separation of reusable behavioral patterns from private details that should not be reproduced.

3.1.2 Third-Party Traces. **Third-party traces** are artifacts produced by others that reveal information about the target person. Examples include peer reviews, performance evaluations, recommendation letters, media reports, customer feedback, code review comments written about the person, and conversations that mention the target individual. For example, a Slack discussion between two teammates may reveal how they perceive a manager's decision-making style, even if the manager is not part of the conversation.

Compared with first-person traces, third-party traces are less directly tied to the person's own behavior. However, they often reveal information that first-person traces cannot easily capture, including reputation, interpersonal impact, blind spots, externally visible competence, and how the person's actions are interpreted by others. Existing person distillation systems rely primarily on first-person traces [109, 135], leaving third-party evidence largely unexplored. Zhangxuefeng-Skill [3] provides a partial example by combining the subject's own books and speeches with analytical commentary from technology and business media.

Research in personality psychology and organizational behavior supports the value of third-party evidence. Connelly and Ones [21] show that other-ratings of personality can predict job performance and academic achievement better than self-ratings and provide predictive power beyond self-reports. Vazire [114] proposes the Self–Other Knowledge Asymmetry (SOKA) model, which argues that individuals are often more accurate in assessing low-observability traits, whereas observers may be more accurate in assessing highly evaluative traits such as intellect. For person distillation,

this suggests that first-person traces may better reveal internal preferences and reasoning processes, while third-party traces may better reveal competence, social impact, and evaluative judgments.

However, third-party traces are not objective ground truth. They reflect the observer's perspective, incentives, relationship with the target person, and social context. A peer review may identify genuine weaknesses but may also reflect bias, misunderstanding, or conflict of interest. A media report may summarize a person's reputation while emphasizing controversy or adopting a particular narrative frame. Consequently, third-party traces should be interpreted as evidence about how a person is perceived rather than direct evidence of their internal state.

3.1.3 Combining First-Person and Third-Party Evidence. First-person and third-party traces provide complementary perspectives on the same individual. First-person traces capture what the person directly said, wrote, built, decided, and experienced. Third-party traces capture how those actions were perceived, evaluated, interpreted, and discussed by others. Together, they provide a more complete representation than either source alone.

For example, a person's architecture decisions may reveal how they reason about technical tradeoffs, while colleagues' assessments may reveal whether those decisions were considered effective, influential, or difficult to work with. Similarly, a person's stated leadership philosophy may differ from how team members experience that leadership in practice. Combining both perspectives therefore offers the potential to capture both internal and external views of the same individual.

This complementarity is consistent with the SOKA framework [114], which argues that different observers possess different kinds of knowledge about a person. Provenance should therefore be treated not merely as metadata, but as a factor that determines which aspects of a person can be reliably inferred from a trace.

Integrating these sources introduces several important research challenges.

Data Integration. First-person and third-party traces are often distributed across different platforms, modalities, and contexts. Integrating them requires identity resolution, event alignment, and provenance-preserving representations that maintain links between observations and their sources.

Verification and Conflict Resolution. Different sources may provide conflicting accounts of the same individual. Such disagreements may arise from self-presentation bias, observer bias, differing perspectives, or genuine changes over time. Future systems must reason about source reliability, uncertainty, and disagreement rather than simply averaging conflicting signals.

Multimodal Provenance Fusion. Provenance interacts with modality. First-person traces may include text, code, audio, video, and behavioral logs, whereas third-party traces may include evaluations, transcripts, social media discussions, and organizational records. Future research must develop methods for combining multimodal evidence across provenance types while preserving the evidential basis of distilled conclusions.

3.2 Trace Modality: How Is the Evidence Expressed?

While provenance concerns *who generated a trace*, **modality** concerns *how the trace is expressed*. Personal traces may appear as text, code, audio, video, images, behavioral logs, relational graphs, and other modalities. Modality matters because different modalities expose different facets of a person. The same individual may reveal expertise through code, communication style through messages, behavioral habits through activity logs, and organizational role through collaboration networks. Thus, the modalities available to a person distillation system determine what aspects of the person can be observed, inferred, and distilled.

Existing systems already use multiple modalities, but they remain concentrated in a narrow region of the design space. Most are multimodal in principle but text-dominant in practice.

3.2.1 Text-Dominant Multimodal Traces. Existing person distillation systems rely heavily on textual artifacts, including chat histories, emails, documents, social media posts, books, interviews, and transcripts [1, 3, 109, 116, 128, 135]. Text is easy to collect, search, index, summarize, and process with modern language models. It is also intentionally produced: people choose what to write, how to phrase it, and what to include. As a result, text often contains concentrated signals about expertise, preferences, communication style, and reasoning patterns.

However, text captures only part of a person. **Audio** reveals how a person communicates, including tone, pacing, hesitation, emphasis, confidence, and emotional expression. These signals often disappear in transcription. **Visual traces** capture other dimensions of behavior. Whiteboard sketches reveal how a person structures problems spatially. Presentation slides show how they organize arguments visually. Recorded demonstrations show how they interact with tools, systems, and collaborators in real time.

Several systems have begun to incorporate such modalities. MamaSkill [42] uses voice memos. Ex-Skill [109] extracts temporal and geographic information from photographs through EXIF metadata. Brother-Skill [1] ingests video, audio, and textual content from YouTube and TikTok. Zhangxuefeng-Skill [3] incorporates speeches and lecture videos.

Despite their value, audio and visual traces remain underused because they are noisy and expensive to process. A one-hour meeting recording may contain only a few minutes of person-revealing behavior. A photograph may contain a useful sketch together with irrelevant background details. This tradeoff between signal richness and processing complexity helps explain why current systems continue to favor text. As multimodal foundation models improve, however, these richer modalities may become increasingly important for capturing aspects of a person that text alone cannot reveal.

3.2.2 Code as a Behavioral Trace. A particularly important but underexplored modality is **code**. Although code is central to agent learning and software engineering research, it plays a limited role in current person distillation systems.

In agent skill learning, code often represents a task solution. For example, Voyager [118] learns reusable skills from generated code, execution feedback, errors, and self-critique. In this setting, code is primarily an executable artifact for accomplishing tasks.

For person distillation, human-authored code should be viewed differently. When a developer structures a module, names variables, designs interfaces, writes tests, or reviews a pull request, they leave traces of architectural thinking, problem decomposition, quality standards, and engineering taste [13]. Code is therefore not merely an artifact of task completion; it is evidence of how a specific person reasons and makes decisions.

Existing work has only begun to explore this opportunity. VibePortrait [23] distills developer characteristics from interactions with AI coding assistants. More broadly, AI-assisted development creates a new class of traces that we call **AI-mediated coding traces**. These include prompts, rejected suggestions, edits, tests, refinements, and acceptance decisions produced during collaboration between a human developer and a coding assistant.

AI-mediated traces occupy an intermediate position between human-authored and machine-generated content. Although many code tokens may originate from the model, the developer frames the problem, specifies constraints, evaluates alternatives, and validates solutions. Thus, the most informative signal may lie not in the final code artifact but in the interaction process: how the developer evaluates suggestions, revises imperfect solutions, and exercises judgment throughout development.

3.2.3 Behavioral Logs and Human Trajectories. Another underexplored modality is **structured behavioral data**. Examples include commit histories, calendar events, task records, interaction timestamps, review activities, communication frequencies, and response latency patterns.

Unlike text, which primarily captures what people say, behavioral traces capture what people do. They record actions, decisions, routines, and temporal patterns. A developer's commit history may reveal when they work, which systems they know best, and how they approach debugging. Calendar records may reveal how they allocate attention across projects. Code review histories may reveal which technical issues they consider important enough to examine closely.

This perspective connects person distillation to agent skill learning. Frameworks such as Trace2Skill [77] and Jiang et al. [41] learn from trajectories consisting of states, actions, and outcomes. Human behavioral traces can be viewed as real-world analogues of such trajectories.

However, human trajectories differ from agent trajectories in a crucial way: they are almost always **incomplete**. Agent systems are usually designed to record observations, actions, and outcomes. Human activity is not. An engineer investigating a production incident may leave behind Slack discussions, dashboard screenshots, commits, and tickets, while critical reasoning occurs in a Zoom meeting, hallway conversation, or private mental process. The available traces are therefore fragments rather than complete trajectories.

This observation motivates a research agenda around **trajectory reconstruction**. Future systems may need to identify task boundaries across platforms, align fragmented observations into coherent episodes, distinguish observed actions from inferred actions, and represent uncertainty arising from missing information. Incomplete trajectories are not a corner case of person distillation; they are the normal case.

3.2.4 Relational Traces and Social Context. A third major gap concerns **relational data**. Existing systems primarily focus on content produced by an individual, while largely ignoring the social structures in which that individual operates.

Relational traces include organizational hierarchies, collaboration networks, communication graphs, code review relationships, mentorship chains, expertise networks, and other representations of how people interact. Such traces describe not the content of a person's actions but their position within a broader social system.

This distinction matters because many aspects of a person's role cannot be inferred from content alone. A senior engineer's Slack messages may reveal technical expertise and communication style. A collaboration graph may reveal that they bridge multiple engineering organizations. A review network may show that junior engineers consistently seek their advice before consulting others. These properties emerge from relationships rather than from any individual message or document.

Current systems occasionally assume relational context. For example, Boss-Skills [116] assumes that the target person is a manager but does not explicitly model relational structure as a first-class source of evidence. This omission limits the ability of person distillation systems to capture how expertise is deployed, how influence propagates, and how a person's role emerges within a larger community.

A key open challenge is **relational representation**. Should relationships be modeled as hierarchies, collaboration graphs, communication networks, review networks, or multi-relational knowledge graphs linking people, topics, projects, and roles? More broadly, relational traces suggest that person distillation should move beyond isolated individuals and model people as participants in evolving social systems.

3.3 Trace Abstraction: How Processed Is the Evidence?

While provenance concerns *who generated a trace* and modality concerns *how a trace is expressed*, **abstraction level** concerns *how much interpretation has already been applied*. Some traces are direct observations of behavior, such as messages, commits, recordings, and activity logs. Others are processed summaries, evaluations, profiles, recommendations, or narratives that compress many observations into higher-level descriptions. This distinction matters because abstraction changes both the information content and the assumptions embedded in the data. As traces become more abstract, they often become easier to store, search, and use, but they also move further away from the original evidence.

3.3.1 Raw Behavioral Traces. At one end of the spectrum are **raw traces**: a person's actual messages, code, meeting recordings, calendar events, behavioral logs, and other direct records of activity. Raw traces are the closest available approximation to observing the person's behavior directly. The person actually sent the message, wrote the code, attended the meeting, or performed the recorded action, with minimal intermediate interpretation.

The primary advantage of raw traces is **fidelity**. Because little processing has been applied, raw traces preserve subtle behavioral patterns that higher-level abstractions may overlook. For example, Goyal et al. [35] show that observing how a person works can predict task performance even without examining the final outcome. Similarly, Stachl et al. [101] demonstrate that passively collected behavioral traces such as mobility patterns, communication activity, and media consumption can predict personality traits. These findings suggest that person-specific signals are often embedded in behavioral details that may appear unimportant in isolation but become informative in aggregate.

Raw traces are especially valuable because they preserve information that was never explicitly articulated. A senior engineer may never state, "I prefer eventual consistency over strong consistency," yet that preference may become visible across many architecture decisions. A manager may never describe their leadership philosophy directly, yet recurring communication patterns, meeting behaviors, and delegation choices may reveal it.

However, fidelity comes at a cost. Raw traces are often noisy, repetitive, and weakly informative when viewed individually. A person's digital history may contain millions of messages, interactions, and events, most of which reveal little about distinctive expertise, judgment, or personality. Processing such traces at scale is expensive, and the relevant signal is often distributed across many observations. Recovering a person's expertise, values, or decision criteria may require aggregating evidence across months or years.

Thus, the central challenge is not merely collecting raw traces but extracting person-specific signal from them without discarding the subtle details that make them valuable.

3.3.2 Interpreted and Derived Traces. At the opposite end of the spectrum are **interpreted traces**: representations that already contain human or algorithmic abstraction. Examples include performance reviews, recommendation letters, personality assessments, resumes, biographies, media profiles, project retrospectives, expert evaluations, and AI-generated summaries of behavior.

Unlike raw traces, interpreted traces do not simply record what happened. They explain, evaluate, or characterize the person. A performance review may summarize a year of work into a few paragraphs. A recommendation letter may compress hundreds of interactions into high-level claims about leadership or technical ability. A manager's statement that an engineer is "excellent at incident response" is itself an abstraction built from many observations.

The primary advantage of interpreted traces is **information density**. Rather than forcing a system to reconstruct patterns from thousands of observations, interpreted traces often expose high-level conclusions directly. They may

identify strengths, weaknesses, expertise areas, or behavioral tendencies, making such information easier to retrieve and use.

However, interpreted traces embed the assumptions, biases, and limitations of whoever produced them. A performance review reflects the reviewer's expectations and criteria. A media profile reflects the author's framing. AI-generated summaries inherit biases from both the underlying data and the summarization process. Consequently, interpreted traces should not be treated as objective ground truth. They are evidence about how behavior has been interpreted, not necessarily direct evidence of the behavior itself.

Moreover, abstraction is lossy. When a year of behavior is compressed into a few sentences, many details disappear. A summary may correctly identify that a person is a strong technical leader while omitting the situations, tradeoffs, and decisions that support that conclusion. As abstraction increases, traces become easier to use but may lose the nuances that distinguish one individual from another.

3.3.3 Linking Raw Evidence and Abstractions. Raw and interpreted traces should be viewed as complementary levels of evidence. Raw traces provide fidelity; interpreted traces provide structure. Together, they form a hierarchy from direct observations to increasingly abstract characterizations.

This hierarchy raises several research challenges.

Evidence Compression. Person distillation requires compressing large volumes of behavioral evidence into representations that can be stored, retrieved, and used efficiently. Excessive compression risks losing person-specific nuance, while insufficient compression leaves systems overwhelmed by raw data.

Traceability and Verification. As abstractions move further from the underlying evidence, verification becomes harder. If a distilled representation claims that a person is an excellent mentor or a risk-averse decision maker, users should be able to trace that claim back to supporting evidence. Future systems may need explicit links between high-level abstractions and the raw traces from which they were derived.

Multi-Level Reasoning. Different downstream tasks require different abstraction levels. Answering a question about how a person handled a specific production incident may require concrete traces. Predicting how that person would approach a future incident may require abstract behavioral patterns learned from many episodes. Person distillation systems may therefore need to reason across multiple abstraction levels, dynamically moving between raw observations and higher-level interpretations.

More broadly, abstraction suggests that person distillation should not construct a single flat representation. It may require a hierarchy of representations that preserves connections among observations, patterns, explanations, and high-level characterizations.

3.4 Trace Temporality: When Was the Evidence Generated?

While provenance concerns *who generated a trace*, modality concerns *how the trace is expressed*, and abstraction level concerns *how much interpretation has already been applied*, **temporality** concerns *when the trace was generated and how its temporal structure is represented*. Temporality matters because people are not static. Their knowledge, preferences, relationships, expertise, responsibilities, and even sense of identity evolve over time [12, 95]. A person distillation system that ignores time risks collapsing multiple versions of a person into a single static portrait.

Longitudinal research shows that personality traits change across the lifespan rather than remaining fixed [89]. Similar evolution occurs in professional settings: individuals acquire expertise, develop judgment, adapt to new organizational

environments, and redefine professional identities throughout their careers [8, 58]. A junior engineer, a senior architect, and an engineering manager may be the same individual at different points in time, yet they may exhibit different behaviors, priorities, and decision-making patterns. Person distillation must therefore account not only for *who a person is* but also for *when they were that person.*

Existing person distillation systems rely on temporally situated traces, including workplace communications, personal conversations, books, speeches, interviews, media reports, and other artifacts [109]. However, temporality is usually treated as metadata rather than as a first-class representational dimension. This is a missed opportunity: temporal structure determines what kinds of change can be observed, what patterns can be reconstructed, and what version of a person is ultimately distilled.

3.4.1 Snapshot and Longitudinal Traces. The most basic temporal distinction is between **snapshot** and **longitudinal** traces. Snapshot traces capture a person at a particular moment or during a limited period. Examples include resumes, profile pages, performance reviews, recommendation letters, interviews, and published articles. Such traces often summarize a longer history into a point-in-time description.

In contrast, **longitudinal traces** preserve observations across extended periods. Examples include commit histories, code review records, communication archives, meeting transcripts, and long-running social media histories. Unlike snapshots, longitudinal traces reveal trajectories. They make it possible to observe how expertise develops, how communication styles change, how relationships evolve, and how behavioral patterns emerge over time [53, 60].

This distinction matters because many person-specific characteristics are inherently temporal. Expertise is not simply possessed; it is acquired. Leadership is not merely a trait; it emerges through repeated interactions. Judgment is not static; it develops through experience. A snapshot may reveal the current state of a person, whereas longitudinal traces reveal how that state was reached.

3.4.2 Facet-Specific Temporal Dynamics. Different facets of a person evolve at different rates. Some facets are largely **cumulative**. Domain expertise, technical knowledge, and professional judgment often develop gradually over years, so older traces may remain informative [58]. An architecture document written five years ago may still reveal enduring principles that guide a person's technical decisions today.

Other facets are more **context-sensitive**. Communication style, collaboration patterns, and organizational roles may change rapidly when a person joins a new team, assumes a new position, or operates under different constraints [43]. A person's communication style as an individual contributor may differ substantially from their communication style as a manager.

Still other facets exhibit **long-term stability with gradual drift**. Personality traits, values, and interpersonal tendencies may remain recognizable over long periods while still evolving through major life experiences and career transitions [71]. These differences suggest that temporality should not be modeled uniformly. The relevance of a trace depends not only on its age but also on the facet being inferred. A decade-old design document may remain valuable evidence of technical judgment, while a decade-old email exchange may be weak evidence of current communication behavior.

3.4.3 Temporal Person Models. Beyond individual traces, temporality raises a deeper representational question: should a person be modeled as a single profile or as an evolving sequence of profiles?

Most current systems implicitly assume a single profile. They aggregate traces across time and distill them into one representation. This approach is simple, but it obscures temporal variation. Contradictions between older and newer

behaviors are often averaged away rather than explained. A person who once preferred rapid experimentation but later became risk-averse may appear inconsistent when, in reality, their priorities changed.

An alternative is to treat person distillation as constructing a **temporal person model**. Rather than producing one static representation, the system would maintain multiple temporally grounded representations connected by an evolution trajectory. Such a model could capture how expertise accumulates, how responsibilities shift, how preferences change, and how major events reshape behavior. Temporal change is therefore not merely noise to filter out; it is part of the person being distilled.

3.4.4 Temporal Challenges. Integrating temporality into person distillation introduces several challenges.

Temporal Scope. What period of a person's life should be included in the source evidence? A short window may better represent the current person while reducing storage and processing costs. A longer history may reveal deeper developmental trajectories but may also incorporate outdated information.

Temporal Attribution. Many traces describe one time period while being created in another. Resumes, interviews, autobiographies, postmortems, and performance reviews often discuss past events from a later perspective. Systems therefore need to distinguish event time from recording time and reason about temporal ambiguity.

Temporal Sparsity and Missing Periods. Personal traces are rarely collected uniformly. Data may be concentrated around particular projects, jobs, life events, or periods of intense activity, while other periods remain unobserved. Future systems may need to model temporal coverage and represent uncertainty arising from missing data rather than assuming continuous observation.

Temporal Drift and Versioning. A fundamental challenge is determining when accumulated changes justify treating the person as a new version of themselves. At what point does a junior engineer become a senior engineer, or an individual contributor become a manager? Future systems may require mechanisms for detecting meaningful shifts in expertise, responsibilities, preferences, or behavior and for maintaining multiple temporally grounded versions of the same person.

More broadly, temporality suggests that person distillation should be viewed not merely as modeling a person but as modeling a person's evolution.

3.5 Gaps in the Source-Data Landscape

The four dimensions above, **provenance**, **modality**, **abstraction level**, and **temporality**, define a broad design space for personal traces. Existing research related to person distillation occupies only small regions of this space. Although several adjacent fields aim to model individuals, characters, users, or agents, they typically focus on narrow subsets of trace types while overlooking the heterogeneous evidence generated by real people.

Two especially relevant lines of work are role-playing LLMs and agent skill distillation. Both provide useful foundations, but neither captures the full source-data complexity required for person distillation.

3.5.1 Role-Playing LLMs: Personas with Limited Evidence. The first related category is **role-playing and persona-based LLMs**, whose goal is to make a model behave consistently as a character, historical figure, or fictional persona. These systems rely on source evidence about individuals, but the evidence is typically narrow, curated, and dominated by conversational content.

At the simplest end, persona-mimicry systems operate almost entirely on dialogue. For example, Li et al. [56] extract dialogues from television and anime scripts to construct memory databases of utterances, while Zhang et al. [132] use crowdworker-generated persona descriptions and conversations to guide role-playing behavior. Similarly, RoleLLM [120] constructs role profiles primarily from dialogues and character descriptions extracted from novels and movies.

More recent systems enrich conversational data with additional context. Character-LLM [98] incorporates information about a character's experiences and emotional states. CoSER [123] further extracts actions and internal thoughts from literary works, either directly from the source text or through LLM-based inference. These additions move beyond pure dialogue, but they remain grounded in narrative descriptions of characters.

Viewed through our taxonomy, role-playing systems occupy a narrow region of the source-data space. Their traces are predominantly textual, curated, and highly abstracted, often organized around scripted interactions. Missing are the heterogeneous, naturally occurring traces that characterize real human behavior: workplace communications, code repositories, meeting recordings, activity logs, behavioral trajectories, collaboration networks, and other modalities through which professional expertise and real-world decision-making become visible.

This limitation becomes clear when considering expertise and judgment. Conversational traces are well suited for capturing how a person speaks, their tone, emotional reactions, and aspects of personality. However, they provide limited visibility into how the person solves problems, evaluates alternatives, and makes decisions under uncertainty. A role-playing system may reproduce how Beethoven speaks in a hypothetical conversation, but such dialogue reveals little about how he composed music: how he evaluated musical structures, departed from convention, or balanced technical constraints against artistic expression. Similarly, a profile stating that someone is an "experienced and meticulous engineer" describes a characteristic but does not expose the reasoning processes that produced that reputation.

From the perspective of person distillation, the key limitation is **evidential depth**. Role-playing systems primarily model how a person appears in conversation. Person distillation seeks to model how a person thinks, decides, collaborates, and develops expertise. Capturing these deeper facets requires traces that extend beyond dialogue and narrative description into the behavioral records of real-world practice.

3.5.2 Agent Skill Distillation: Behavior without Identity. The second related category is **agent skill distillation**, whose objective is to extract reusable skills from successful agent behavior. Rather than dialogues and persona descriptions, agent skill frameworks rely on task execution trajectories.

For example, Voyager [118] learns reusable skills from generated code, environment feedback, execution errors, and self-reflection. Trace2Skill [77] distills skills from trajectories consisting of task descriptions, reasoning traces, tool-use histories, outputs, and correctness signals. More broadly, Jiang et al. [41] conceptualize skill distillation as a lifecycle centered on task discovery, refinement, execution, and evaluation.

Viewed through our taxonomy, these systems occupy a different region of the source-data space. They use structured behavioral traces, preserve temporal trajectories, and focus on observable actions and outcomes. In many ways, they are closer to person distillation than role-playing systems because they model behavior rather than description.

However, they pursue a different objective. Agent skill distillation extracts **task-general capabilities**. Its central question is whether the agent successfully completed the task and what reusable procedure can be learned from that success. The identity of the actor is largely irrelevant.

Person distillation asks a different question. Rather than asking, "How can this task be solved?", it asks, "How would this particular person solve it?" This distinction matters when multiple individuals achieve the same outcome through

different reasoning processes. Two senior engineers may both resolve the same production outage: one by systematically narrowing the search space through hypothesis testing, the other by relying on years of accumulated intuition. From the perspective of agent skill distillation, both trajectories demonstrate a successful debugging skill. From the perspective of person distillation, the difference is precisely the signal of interest.

More broadly, person distillation requires forms of evidence that agent skill frameworks typically ignore. Communication traces reveal how individuals explain and justify decisions. Code review comments reveal quality standards and evaluative criteria. Meeting recordings reveal persuasion strategies and stakeholder management. Relational traces reveal mentorship, influence, and organizational roles. These forms of evidence may contribute little to task success, yet they are central to understanding what makes one individual distinct from another.

Consequently, agent skill distillation captures **behavior without identity**, while role-playing systems capture **identity without behavior**. Person distillation sits at the intersection of these traditions. It requires the behavioral richness of agent trajectories together with the person-specific characteristics that make those behaviors uniquely attributable to a particular individual.

Existing systems either model how a person *appears* or how a task is *performed*. Person distillation requires modeling how a particular individual acquires knowledge, exercises judgment, communicates, collaborates, and acts across diverse contexts. Achieving this goal requires expanding beyond the narrow regions of the source-data taxonomy currently explored by adjacent fields and embracing the full diversity of traces that real people generate throughout their professional and personal lives.

4 Distilled Person Representations

Section 3 characterized the input to person distillation: the diverse personal traces from which a person's knowledge, behavior, and identity can be inferred. We now turn to the output. *Once information has been distilled from these traces, how should it be represented?* The answer determines what aspects of the person are preserved, how the representation can be inspected, updated, and shared, and ultimately what downstream tasks it can support. We refer to these outputs collectively as **distilled person representations**.

We organize the representation design space along two orthogonal dimensions. The first concerns **what** should be represented: which facets of a person should be preserved, including domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context (Section 4.1). The second concerns **how** those facets should be represented, including document-based representations, parametric models, preference and reward models, and memory and knowledge systems (Section 4.2). Jointly, these two dimensions define a representation design space for person distillation. Figure 4 summarizes this two-dimensional representation design space.

A central observation of this section is that *no single representation format is well suited to every facet of a person.* Explicit knowledge may be naturally expressed in documents, recurring behavioral patterns may be better captured by learned models, judgments may be encoded as preference functions, and episodic experiences may be preserved in memory systems. The challenge is therefore not to identify one universal representation, but to understand how different representation formats complement one another and how they can be combined into a coherent person representation. We first introduce the two taxonomies independently and then discuss how they interact, highlighting the opportunities and challenges of hybrid representations (Section 4.3).

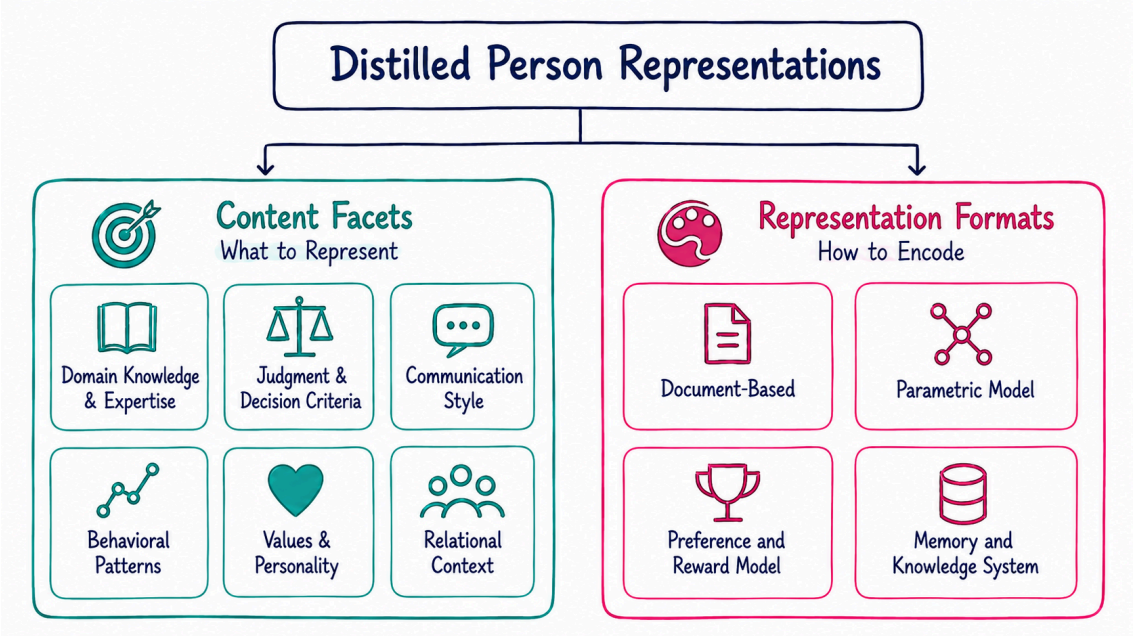

Fig. 4. Taxonomy of the representation design space. Distilled person representations are organized along two dimensions: the content facet dimension, which specifies what facets of a person should be preserved, and the representation format dimension, which specifies how those facets should be represented.

4.1 Content Facets: What to Represent

The first dimension of the representation design space concerns *what* should be represented. This question is more fundamental than choosing a representation format: before deciding how to encode a person, we must first decide **which aspects of the person are worth preserving**. A person is not a single homogeneous object. Domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context capture different aspects of human cognition and interaction, evolve at different rates, and serve different downstream purposes. A faithful person representation should therefore preserve not only *what* a person knows, but also *how* the person thinks, communicates, acts, and relates to others.

This perspective distinguishes person distillation from adjacent research areas. Role-playing LLMs primarily represent communication style and personality [56, 120], whereas agent skill frameworks primarily represent reusable task-solving procedures [77, 118]. Neither attempts to represent the full spectrum of person-specific characteristics. We therefore organize the content of a distilled person into the following six complementary facets.

(1) **Domain knowledge and expertise** (*what they know*) [28]. This facet captures a person's accumulated knowledge, conceptual frameworks, technical expertise, and professional experience. For example, a senior backend engineer may know that PostgreSQL advisory locks provide an effective solution for distributed rate limiting without requiring external infrastructure.

(2) **Judgment and decision criteria** (*how they evaluate*) [44]. This facet captures how a person compares alternatives, balances tradeoffs, and applies quality standards when making decisions. Two engineers may possess

similar knowledge of distributed systems, yet one consistently prioritizes consistency while another accepts eventual consistency in exchange for lower latency.

(3) **Communication style** (*how they express*) [83]. This facet captures characteristic patterns of communication, including tone, wording, argumentation, level of detail, and adaptation to different audiences. For example, one engineer may write code review comments as direct imperatives, whereas another conveys the same feedback through questions and suggestions.

(4) **Behavioral patterns** (*what they do*) [73]. This facet captures recurring behaviors across situations rather than isolated actions. For example, during a production incident, one engineer may immediately inspect dashboards and logs, whereas another first gathers teammates to discuss hypotheses before examining system state.

(5) **Values and personality** (*who they are*) [67]. This facet captures relatively stable dispositions, including personality traits, values, interpersonal orientation, and risk tolerance. For example, one person may naturally think aloud and seek early feedback, whereas another prefers to reflect privately before sharing ideas.

(6) **Relational context** (*how they interact with others*) [4]. This facet captures interaction patterns that depend on particular people and relational contexts. The same engineer may communicate formally with senior leadership, casually with close teammates, and adopt different mentoring styles for different junior engineers.

These six facets are **complementary rather than interchangeable**. Domain knowledge describes *what* a person knows, whereas judgment determines *how* that knowledge is applied. Communication style describes how decisions are expressed, whereas behavioral patterns describe how they are enacted. Values and personality capture relatively stable dispositions, whereas relational context captures how those dispositions are adapted across people and situations. Together, the six facets provide a structured decomposition of the information that distinguishes one individual from another.

More importantly, this taxonomy suggests that **a person should not be represented as a single latent vector or profile**. Different facets exhibit fundamentally different statistical and semantic properties. Knowledge accumulates, judgment matures through experience, communication adapts to audience, behavior depends on context, personality evolves slowly, and relational context depends on the people involved. Treating these heterogeneous properties as one undifferentiated representation risks losing precisely the distinctions that make a person unique.

This perspective also clarifies the limitations of existing paradigms. Role-playing systems primarily capture communication style and personality, while agent skill frameworks emphasize domain knowledge and executable procedures. Neither provides comprehensive coverage of all six facets. Person distillation therefore requires integrating multiple complementary facets rather than extending any single existing paradigm.

Finally, the taxonomy raises a broader research question: **which facets should remain person-specific, and which can be shared across individuals?** Two engineers may possess similar expertise but apply different judgment; two managers may communicate in similar ways while exhibiting fundamentally different leadership styles. Future person-distillation systems should therefore reason about similarity at the level of individual facets rather than treating every person as an indivisible representation. Such a *factorized view of person representations* opens the possibility of transferring, comparing, updating, and evaluating different aspects of a person independently while preserving their overall identity.

4.2 Representation Formats: How to Encode

Having identified what a distilled person should preserve, we now ask how such information should be encoded. We distinguish four representation formats: document-based representations, parametric model representations, preference and reward model representations, and memory and knowledge system representations. These formats should not be viewed as competing alternatives. Rather, they represent different answers to a fundamental question: **What is a distilled person?** Should the person be represented as something that can be *read*, *imitated*, *evaluated*, or *remembered*? Each representation makes different tradeoffs among transparency, learnability, auditability, updatability, privacy, and generalization.

Running example. Throughout this section, consider a senior software engineer who is leaving a company after eight years. The company hopes to build an AI assistant that preserves the engineer's knowledge, judgment, communication style, and experience so that future employees can consult it. The same engineer can be represented in fundamentally different ways depending on the chosen representation format. We use this example to illustrate the four representations below.

4.2.1 Document-Based Representations. Document-based representations encode a distilled person as structured text, such as Markdown skill files, persona cards, system prompts, decision heuristics, or role profiles. This is the dominant representation format in current practice. Colleague-Skill packages distilled knowledge as a Markdown "capability track + bounded behavior track" [135]. Role-playing systems similarly represent characters through textual persona descriptions, catchphrases, and structured dialogue examples [120, 132]. In the running example, the departing engineer is represented by a Markdown profile summarizing areas of expertise, decision heuristics, communication principles, and frequently used engineering practices. Future employees consult this document whenever they seek advice.

The defining advantage of document-based representations is **explicitness**. A document can be read, edited, versioned, shared, audited, and transferred across LLM systems without retraining. It also supports direct correction: when a user says that the distilled agent "would not do that," the corresponding rule or behavioral description can be revised. Document-based representations are therefore naturally suited to explicit knowledge, stated preferences, behavioral guidelines, and user-facing inspection.

Their central limitation is that they treat a person as a *specification*. Not everything important about a person is specifiable. Much expertise is tacit, contextual, and difficult to articulate [78]. Even when relevant knowledge can be written down, the document is useful only if the downstream model can interpret and apply it. Character-LLM shows that models may still hallucinate beyond a character's knowledge boundary despite carefully constructed profiles [98]. Moreover, current practice typically feeds a Markdown skill file directly into an LLM agent [109, 135], conflating three distinct concerns: the quality of the representation itself, the model's interpretation of that representation, and the model's ability to execute the represented behavior. This conflation makes it difficult to determine whether failures originate from the representation, the downstream model, or both.

4.2.2 Parametric Model Representations. Parametric representations encode person-specific patterns directly in model parameters or embeddings rather than external documents. This can be achieved through full fine-tuning, lightweight adapters such as LoRA [37], quantized adapters such as QLoRA [25], or learned persona embeddings. Instead of reading an explicit profile, the model itself is modified or conditioned to behave like the target person.

Existing work has explored this representation primarily in role-playing and personalization. CoSER trains role-playing models from given-circumstance acting data [123]. Character-LLM fine-tunes character-specific models on

reconstructed, profile-grounded scenes [98]. Persona-Plug represents users through learned personal embeddings derived from historical behaviors [61]. These systems demonstrate that person-specific behavior can be encoded within model parameters, although current methods mainly focus on conversational style, character consistency, or user preference rather than the broader facets of person distillation. In the running example, the engineer is no longer represented by a profile. Instead, the model is fine-tuned on years of emails, code reviews, design documents, and technical discussions so that these person-specific patterns are absorbed directly into the model.

The defining strength of parametric representations is **implicit learning**. They can capture recurring patterns that are difficult or impossible to describe explicitly, such as phrasing habits, stylistic preferences, recurring behavioral tendencies, or subtle decision heuristics. Rather than requiring every important characteristic to be manually documented, the representation learns statistical regularities directly from personal traces.

Their central limitation is **opacity**. Unlike a document, a fine-tuned model or adapter cannot be directly inspected to determine what has been learned. Fine-tuning may also capture surface imitation more readily than underlying competence: a model may reproduce vocabulary, tone, and repeated phrases while failing to recover the reasoning processes that produced them. Parametric representations further introduce deployment and privacy challenges. Adapters are often tied to specific foundation models, require retraining after model upgrades, and may memorize sensitive personal information that is difficult to locate or remove.

4.2.3 Preference and Reward Model Representations. Preference and reward model representations encode a distilled person as an evaluator rather than a generator. Instead of producing what the person would say or do, the model scores, ranks, or critiques candidate outputs according to what the person would approve, reject, or consider high quality [10, 57, 100]. This representation is particularly natural for modeling judgment, where the objective is to determine which of several alternatives best reflects the person's standards.

This representation builds upon reward modeling and reinforcement learning from human feedback, where reward models are learned from human comparisons and subsequently guide generation [7, 20, 80]. Recent work extends these ideas toward personalized preference modeling. LoRe learns user-specific reward functions through low-rank structure [10]; PReF represents individual rewards as combinations of shared basis reward functions [100]; and PersRM-R1 develops personalized reward models using synthetic preference data together with supervised fine-tuning [57]. In the running example, the engineer is represented as a reviewer rather than a speaker. Given several candidate code reviews or architecture proposals, the reward model ranks them according to which one the engineer would most likely approve.

The defining contribution of reward models is the **separation of judgment from generation**. A person-specific reward model can evaluate alternatives, guide optimization, support delegation when the person is unavailable, and potentially combine multiple evaluative perspectives. In this view, the distilled person is represented primarily as a source of standards rather than a source of actions.

The limitation is that *evaluation is not creation*. A reward model can only assess the candidates it receives. If none reflects the solution the person would have proposed, the model can only select the closest approximation. Moreover, a scalar reward often hides the structure of the person's reasoning: a low score may arise from poor technical quality, insufficient evidence, inappropriate tone, a value conflict, or a violation of the person's knowledge boundary. More fundamentally, reward models may learn generic notions of quality rather than the person's distinctive judgment. Person distillation therefore requires reward models that capture not only what is good, but *why this particular person considers it good.*

4.2.4 Memory and Knowledge System Representations. Memory and knowledge system representations encode a distilled person as a queryable collection of organized evidence rather than as a compressed profile or model checkpoint. Instead of directly representing the person, they organize retrievable episodes, semantic facts, procedural knowledge, relationships, timestamps, confidence estimates, and links to supporting traces. Agent memory systems such as MemGPT [81] and Mem0 [19], together with cognitive architectures such as CoALA [104], provide useful infrastructure for this representation.

Importantly, the memory representation is not the raw trace archive itself. Emails, Slack messages, code reviews, meeting transcripts, and documents remain source traces. The memory layer organizes these traces into structured representations. In the running example, design documents, incident reports, code reviews, and Slack discussions are converted into retrievable episodes, semantic facts, and procedural memories. When a future engineer asks how the departing engineer would approach a production issue, the system first retrieves relevant experiences before generating an answer.

The defining advantage of memory-based representations is **traceability**. Every high-level conclusion can, in principle, be linked back to supporting evidence. Memory systems also naturally support continual growth and temporal evolution: new experiences can be incorporated, obsolete memories updated or forgotten, and incomplete trajectories represented without pretending that every reasoning step was observed.

The limitation is that *memory is not reasoning*. A memory system is useful only if it retrieves the right evidence at the right time. Poor retrieval may produce convincing answers grounded in irrelevant traces, while successful retrieval does not automatically imply successful generalization to new situations. Memory systems must additionally decide what to store, merge, summarize, forget, protect, and expose. Consequently, memory representations preserve evidence exceptionally well but require sophisticated mechanisms for retrieval, abstraction, temporal reasoning, and privacy management.

Taken together, the running example illustrates that the *same* engineer admits four very different computational representations: a document that can be *read*, a model that can *imitate*, a reward function that can *evaluate*, and a memory system that can *remember*. This observation leads to a broader conclusion: **person representation is inherently multidimensional**. No single representation format is sufficient in isolation. A mature person-distillation system will likely integrate these complementary representations, allowing each to capture the facets for which it is best suited while jointly forming a coherent, inspectable, and continually evolving model of a person.

4.3 Connecting Content Facets and Representation Formats

The two taxonomies above should not be interpreted independently. The content facets define *what* distinguishes one person from another, whereas the representation formats define *how* those characteristics can be represented. The central question is therefore not *which representation is best*, but **how different representations should work together**. Figure 5 illustrates this mapping using a running example. Different facets produce different kinds of evidence, evolve at different rates, and support different downstream tasks. Consequently, no single representation is sufficient for faithfully modeling a person.

The running example illustrates this naturally. Suppose the company has successfully built an AI assistant representing the departing senior engineer. Which representation should answer a user's question? If a new engineer asks, "What systems did she build?", the answer should come from explicit documents or verified memories. If the question is "How would she review this pull request?", the system should rely on learned behavioral patterns and evaluative judgment. If the user asks why the answer is trustworthy, the system should retrieve supporting evidence from memory. The same

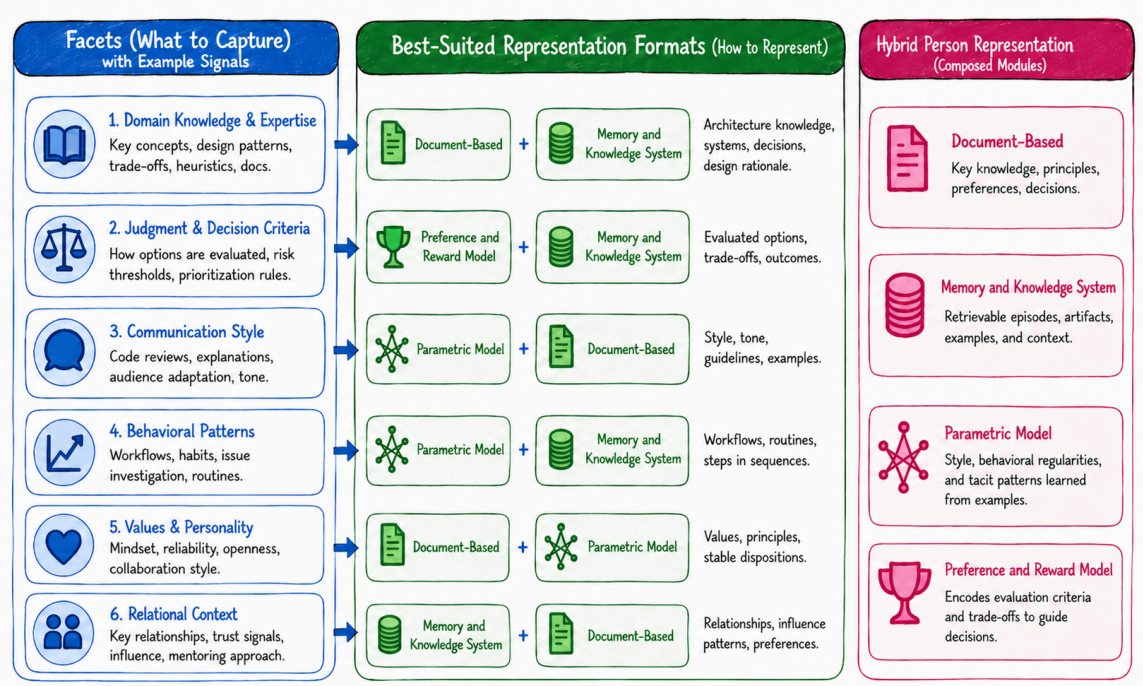

Fig. 5. Mapping from content facets to hybrid representations using the running example of a departing senior software engineer. Different facets are supported by different representation formats, which are composed into a hybrid person representation.

question may therefore require several representations working together rather than one representation operating alone.

This observation suggests the first design principle: **different facets require different representations**. Explicit knowledge, authored documents, and stated preferences are naturally represented by documents and memories. In contrast, tacit expertise often emerges only through repeated decisions and accumulated experience [78, 85]. Likewise, communication style and behavioral regularities are more naturally learned from repeated examples than explicitly described. Reward models further capture what the person consistently approves or rejects [7, 10, 20, 57, 80, 100]. Future person-distillation systems should therefore distinguish between *knowledge that can be stated* and *knowledge that must be learned*.

The second design principle is that **evidence should never be separated from conclusions**. Suppose the assistant recommends rejecting a system design because it lacks sufficient observability. A user should be able to ask, "Why do you think she would say that?" and receive not only an explanation but also the historical code reviews, design discussions, or incident reports that support the conclusion. Memory systems naturally provide this evidential grounding [19, 81, 104], while reward models and document-based summaries become substantially more trustworthy when they remain linked to their supporting traces. This suggests that **traceability** should be treated as a first-class design objective rather than an implementation detail.

The third design principle is that **different facets should evolve independently**. Returning to the running example, the engineer's technical knowledge may continue growing, communication style may gradually change after

becoming a manager, and relationships with teammates may evolve rapidly, while core values remain relatively stable. These heterogeneous dynamics imply that different representations should also be updated in different ways. Stable characteristics may gradually evolve through parametric learning; rapidly changing information may remain in external memories; and evaluative criteria may be refined through continual preference learning. Future person-distillation systems should therefore support *facet-specific evolution* instead of periodically rebuilding an entire representation from scratch.

These principles also explain why existing paradigms capture only part of the problem. Role-playing systems represent communication style and personality through textual descriptions or conversational fine-tuning [56, 110, 120, 132], but provide limited support for grounded expertise, evolving judgment, or relationship-aware behavior. Agent skill frameworks represent executable procedures and behavioral policies [77, 118], but largely ignore person-specific communication, values, and relational context. Neither paradigm alone captures the multidimensional nature of a real person.

Ultimately, **person representation is not a modeling problem but a systems problem**. The AI assistant in our running example is not a document, a model checkpoint, a reward function, or a memory database. It is a coordinated system in which documents preserve explicit knowledge, parametric models capture implicit regularities, reward models represent evaluative judgment, and memory systems preserve grounded evidence. The next generation of person-distillation systems will therefore be defined not by inventing a single better representation, but by learning how to orchestrate multiple complementary representations into a coherent, continually evolving, inspectable, and trustworthy computational model of a person.

5 Distillation Methods

Sections 3 and 4 characterized the two endpoints of person distillation: personal traces as input and distilled person representations as output. This section focuses on the transformation between them. Specifically, we ask how heterogeneous traces, such as messages, documents, code reviews, behavioral logs, meeting records, and peer feedback, can be distilled into faithful computational representations of a particular person.

Current person-distillation systems largely adopt a prompt-based trace-to-profile paradigm. They collect personal traces and use an LLM to generate a Markdown profile, skill file, persona description, or behavioral summary [109, 135]. This approach is attractive because it is simple, transparent, editable, and easy to integrate into existing LLM agents. However, it is also fundamentally compressive. Summarizing heterogeneous traces into a single profile inevitably loses information, as observed in summarization and context-compression settings [86]. The resulting representation also depends heavily on prompt design [133] and may fail to capture tacit, distributional person-specific patterns that are difficult to express explicitly [78]. Consequently, prompt-based methods may produce a plausible persona while failing to preserve the distinctive knowledge, judgment, and behavior of the target person.

Fortunately, adjacent research areas provide many useful methodological foundations. Role-playing and personalized LLMs study persona construction, fine-tuning, learned embeddings, and persona-conditioned generation [94, 98, 120, 132, 136]. Agent skill distillation learns reusable procedures from task trajectories, execution feedback, and tool-use histories [41, 77, 118]. Preference and reward modeling learns evaluative functions from human choices [7, 10, 20, 57, 80, 100]. Agent memory systems organize long-term experience into retrievable records [19, 81, 104]. These methods provide valuable building blocks, but they optimize objectives such as character consistency, task success, or user preference rather than fidelity to a particular real person. Person distillation therefore requires adapting these ideas to preserve evidence grounding, person specificity, multi-facet fidelity, boundary awareness, and continual revision.

The remainder of this section organizes the methodology landscape into six complementary families. We begin with prompt-based trace-to-profile distillation as the current practical baseline, and then discuss memory-based, parametric and adapter-based, preference- and reward-based, behavioral and trajectory-based, and hybrid and revisable methods as increasingly expressive mechanisms for preserving person-specific information. Rather than surveying these methods in isolation, we examine what each family contributes to person distillation, where it falls short, and what methodological advances are needed to build faithful, grounded, and revisable representations of real people.

5.1 Prompt-Based Trace-to-Profile Distillation

Prompt-based trace-to-profile distillation is the current practical baseline for person distillation. A system collects personal traces, provides selected traces or summaries to an LLM, and asks it to generate a structured profile, Markdown skill file, persona description, decision guide, or behavioral summary. Colleague-Skill, for example, packages workplace knowledge into Markdown-style skill representations [135], while Ex-Skill constructs similar representations from personal communication logs and related artifacts [109]. Related role-playing systems also rely heavily on textual profiles, persona descriptions, catchphrases, and dialogue examples [98, 120, 132].

The popularity of this approach stems from its simplicity. The resulting profile is transparent, editable, and portable across different LLM agents without retraining. A user can directly inspect the distilled knowledge, revise incorrect descriptions, or incorporate the profile into downstream applications. These advantages make prompt-based distillation an attractive starting point for practical systems.

Its main limitation is that it treats a person as a written specification. Compressing heterogeneous traces into a single profile inevitably loses information, particularly tacit knowledge and behavioral regularities that are difficult to describe explicitly [78]. For example, two engineers may both be summarized as "cautious about reliability," yet one consistently focuses on observability while the other emphasizes ownership and operational responsibility. Although the profiles appear similar, the underlying judgment criteria are fundamentally different. Furthermore, the generated profile is not an executable policy. A downstream LLM must still interpret and apply the profile, making the final behavior sensitive to prompt wording and context [93, 133]. Because profile statements are often disconnected from their supporting traces, they are also difficult to verify, revise, or audit. Finally, a profile may encourage the model to answer questions beyond the available evidence, leading to unsupported personalization. Character-LLM reports a related phenomenon in which models hallucinate beyond a character's intended knowledge boundary despite carefully constructed profiles [98].

These limitations suggest that prompt-based trace-to-profile distillation should be viewed as structured knowledge extraction rather than ordinary summarization. We argue that future systems should satisfy three requirements.

- **Evidence-linked profile generation** connects every person-specific claim to supporting traces. Instead of producing a free-form profile, the system generates structured claims together with supporting evidence, timestamps, provenance, and confidence estimates. The resulting profile remains auditable and can be automatically verified as new evidence becomes available.
- **Facet-aware profile generation** organizes the profile according to person content facets rather than a single persona description. Claims about domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context are generated separately before being integrated into a unified profile, preventing distinctive characteristics from being collapsed into coarse role-level descriptions.

- **Boundary-aware profile generation** explicitly represents uncertainty and unsupported areas. Rather than encouraging the downstream model to infer missing preferences or knowledge, the profile records what is unknown, weakly supported, or outside the available evidence, enabling the system to abstain when appropriate.

Prompt-based distillation remains an attractive baseline because of its simplicity and interpretability. However, future work should move beyond profile generation by developing loss-aware compression, incorporating evidence metadata and confidence estimates, measuring facet coverage, checking consistency across profile components, and evaluating profiles against strong matched-peer baselines. These directions would transform prompt-based distillation from an informal summarization technique into a more faithful, evidence-grounded, and auditable methodology.

5.2 Memory-Based Distillation

Memory-based distillation extends the prompt-based paradigm by replacing a single global profile with a collection of *distilled memory units*. Rather than summarizing all personal traces into one document, the system organizes person-specific evidence into retrievable memories. This approach preserves substantially more information while maintaining explicit links to supporting evidence. Existing person-distillation systems primarily rely on document-based profiles [109, 135], whereas retrieval-augmented generation (RAG) [55], long-term agent memory [19, 81, 82], and related memory architectures suggest a richer alternative for representing people.

The motivation follows directly from the limitations of prompt-based profiles. A single profile inevitably loses case-specific evidence, rare exceptions, and context-dependent behavior. Memory-based distillation instead preserves these signals as individual memory units. For example, design decisions, code reviews, production incidents, mentoring interactions, and relationship-specific communication patterns can each become retrievable memories. When asked how a person would approach a new problem, the system retrieves the most relevant experiences before generating a response, transforming person representation from a static profile into an evidence-grounded memory system.

A practical memory-based pipeline consists of two stages: *memory construction* and *person-centered indexing*. The first stage builds directly upon existing work in retrieval-augmented generation, long-term agent memory, and process mining [19, 55, 81, 82, 111]. Raw traces are segmented into memory candidates, related fragments are consolidated into coherent episodes, and each memory stores metadata such as provenance, timestamps, participants, source artifacts, and uncertainty. This produces a persistent collection of structured memories rather than a single compressed profile.

However, directly applying RAG or agent memory is insufficient for person distillation. Conventional retrieval primarily retrieves documents that are topically similar to a query. Person distillation requires retrieving evidence that explains how a particular person would think, judge, communicate, or act. A past code review, for example, may provide the best evidence for a person's reliability standard even if it concerns a different service than the current design proposal. Consequently, memories should be indexed not only by topic, but also by person-specific facets such as domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context. Retrieval therefore becomes person-centered rather than document-centered.

This perspective naturally raises the problem of *facet assignment*. Each memory should be associated with the person content facets it supports. A code review comment may simultaneously provide evidence of technical judgment and communication style, whereas a meeting transcript may primarily reveal relational context. Future work could investigate supervised classifiers, LLM-based judges, and hybrid human–LLM annotation protocols for assigning memories to person content facets while controlling annotation reliability [62, 115].

Compared with prompt-based distillation, memory-based distillation provides stronger evidence grounding, traceability, and support for continual updates. New experiences can be incorporated as new memories, outdated memories can be revised or deprecated, and conflicting evidence can be preserved instead of being averaged into a single summary. At the same time, memory alone is not sufficient. A memory system primarily stores evidence; it does not automatically infer general decision rules, abstract recurring behavioral patterns, or generalize to novel situations. Consequently, memory-based distillation is naturally complementary to prompt-based, parametric, and other learning-based approaches.

Several research challenges remain. First, person-centered indexing should retrieve memories according to person-specific reasoning rather than topical similarity. Second, retrieval itself should be learned so that retrieved memories maximize prediction of the target person's behavior instead of semantic relevance alone. Third, memory systems should support abstraction with traceability by summarizing repeated experiences while preserving links to the underlying evidence. Fourth, they should explicitly represent conflicting evidence and its provenance instead of collapsing disagreements into a single memory. Finally, retrieval should support principled abstention: when no supporting evidence exists, the system should acknowledge this rather than fabricate person-specific behavior. These directions move memory-based distillation beyond conventional RAG toward an evidence-grounded representation of a real person.

5.3 Parametric and Adapter-Based Distillation

Prompt-based methods distill a person into a single textual profile, while memory-based methods preserve a collection of evidence-backed memories. Parametric and adapter-based methods represent a third alternative: they distill person-specific information directly into learned model parameters. Instead of explicitly storing what a person knows or has done, the model learns recurring behavioral patterns from many examples. Although current person-distillation systems mainly rely on external artifacts such as Markdown profiles or skill files [109, 135], adjacent research already demonstrates the effectiveness of fine-tuning, parameter-efficient adapters, learned embeddings, and conditioning modules for adapting LLM behavior, including LoRA [37], QLoRA [25], role-playing models [98, 120], and personalized LLMs [94, 136].

The motivation is that many person-specific characteristics are distributional rather than explicit. A person may never state "I prefer designs with strong observability," yet this preference may consistently appear across years of code reviews and incident responses. Likewise, a manager's feedback style may emerge only through repeated interactions rather than explicit rules. Such recurring regularities are difficult to summarize in a profile and cumbersome to recover from individual memories. Parametric representations instead learn these patterns directly from examples and use them to guide generation, ranking, or action selection.

A typical pipeline converts personal traces into supervised training instances. Communication traces become examples of writing style, review histories become examples of technical judgment, historical choices become preference signals, and behavioral trajectories become supervision for action prediction. These examples are then used to train person-specific adapters, learned embeddings, or other lightweight parameterizations that condition the underlying language model.

However, directly adopting role-playing or personalization methods is insufficient for person distillation. Role-playing models primarily learn conversational style and character consistency [98], while personalized LLMs often optimize task-specific user preferences such as writing style, tagging, ratings, or citation prediction [94]. Person distillation requires substantially richer fidelity, including domain knowledge and expertise, judgment and decision criteria, communication

style, behavioral patterns, values and personality, and relational context. A model that merely sounds like the target person or remembers historical details is therefore not a faithful representation.

Parametric representations also introduce several challenges. Because person-specific knowledge is encoded implicitly, the learned representation is difficult to inspect [26], difficult to edit when mistakes are discovered [70], and vulnerable to memorizing sensitive personal information [15]. Moreover, learned adapters may overfit superficial stylistic cues while failing to capture deeper reasoning patterns [27]. Unlike memory-based methods, they also provide little direct evidence for why a particular output reflects the target person, making grounding and boundary awareness significantly more difficult.

These challenges suggest that person distillation should move beyond a single monolithic adapter. Instead, we propose *facet-specific* and *evidence-regularized* parametric representations. Different adapters or learned modules can specialize in domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, or relational context, while their outputs remain grounded in supporting memories or profile claims whenever possible. Rather than replacing prompt-based or memory-based representations, parametric models should complement them by learning the recurring regularities that are difficult to encode explicitly.

This perspective raises several methodological questions. First, traces must be assigned to the facets they supervise. Because individual traces often support multiple facets, routing should naturally be multi-label. Practical solutions may combine limited human annotation, active learning [97], and LLM-based judges with explicit rubrics [62, 115], and report annotation reliability [6]. Second, multiple facet-specific modules must be composed at inference time. Possible directions include router-based composition inspired by adapter fusion and mixture-of-experts architectures [29, 84, 99], or constraint-based composition where separate modules govern judgment, communication style, and boundary checking. Third, conflicts among modules must be resolved, such as disagreement between communication style and judgment. Finally, local model editing remains an open problem. Although recent model-editing methods provide useful foundations [70, 74], person distillation requires edits that are facet-specific, evidence-grounded, and aligned with feedback from the target person.

Future work should therefore investigate facet-specific training objectives, reliable local model editing, privacy auditing of learned parameters, tighter integration with memory-based retrieval, and evaluation against matched-peer baselines. The central question is not whether learned parameters imitate a person's style, but whether they capture person-specific knowledge and reasoning beyond what can be achieved with explicit profiles and retrieved memories while remaining grounded, corrigible, and trustworthy.

5.4 Preference- and Reward-Based Distillation

Prompt-based methods distill what a person says into an explicit profile, memory-based methods preserve what the person has experienced as retrievable memories, and parametric and adapter-based methods learn recurring behavioral patterns from large collections of traces. Preference- and reward-based distillation focuses on a complementary person facet: *judgment and decision criteria*. Instead of modeling what a person would produce, it models how the person evaluates alternatives. Given a context and a set of candidate responses, the learned preference model predicts which alternatives the target person would prefer or reject. The resulting model can rank candidates, critique outputs, select actions, or guide generation toward decisions that better reflect the person's evaluative standards.

The central challenge is that a person's preferences are rarely observed explicitly. In standard preference learning and reinforcement learning from human feedback (RLHF), supervision is typically provided through ratings, rankings, or pairwise comparisons, which become direct training signals for reward models [20, 80, 103]. In person distillation,

however, evaluative standards are usually implicit. They must be inferred from traces such as edits, comments, revisions, and repeated choices. The goal is therefore not only to determine which alternative the person preferred, but also to recover the underlying judgment criteria and identify the situations in which those criteria should generalize.

Three design questions arise naturally.

The first question concerns *preference scope*. A person's judgment often depends on context, including the task domain, the artifact being evaluated, the intended audience, the person's role, the stakes, and the available evidence. Consequently, a single global preference model is unlikely to capture the full range of a person's evaluative behavior. Preference- and reward-based methods may instead be scoped in several ways. A top-down strategy defines models for specific application domains, such as code review, writing revision, or strategic planning. A facet-based strategy organizes models around different kinds of judgment, such as technical correctness, maintainability, communication tone, or relational appropriateness. An evidence-driven strategy groups traces exhibiting similar contexts, artifacts, or decision patterns. Existing work on fine-grained and multi-objective reward modeling suggests decomposing holistic rewards into more specific evaluative dimensions [119, 125], while mixture-of-experts reward models provide one possible architecture for combining heterogeneous preference criteria [129]. These ideas suggest representing a person through multiple context-specific preference or reward models rather than a single global reward function.

The second question concerns *preference supervision*. Some traces provide direct preference evidence, including ratings, rankings, or pairwise comparisons [20, 80, 103]. More commonly, supervision must be constructed from observed behavior. Useful comparison pairs may come from accepted versus rejected proposals, approved versus declined requests, or competing alternatives discussed during decision making. Preference information can also be extracted from revisions, such as before-and-after edits [63, 69], or constructed from model-generated alternatives [80, 103], counterfactual edits [47], or matched-peer alternatives [10, 16]. These comparisons provide indirect evidence of a person's evaluative standards even when explicit preference labels are unavailable.

The third question concerns *preference representation*. Comparison data can be used to train an explicit reward model [20, 80, 103], directly optimize a generator through preference learning [87], learn personalized or factorized reward functions [10, 16, 100], or model multiple judgment dimensions simultaneously [125]. Each representation makes different tradeoffs between interpretability, sample efficiency, and generalization. For person distillation, however, the learned model should preserve not only which alternative is preferred, but also why the preference is specific to the target person, where it applies, and when the available evidence is insufficient to support a confident judgment.

Compared with the previous three method families, preference- and reward-based methods naturally capture evaluative reasoning but provide weaker evidence grounding. A reward score indicates which alternative is preferred, but not necessarily which past experiences or observations justify that judgment. Preference- and reward-based methods are therefore best viewed as complementary to prompt-based profiles, memory-based representations, and parametric models. Profiles summarize explicit principles, memories provide supporting evidence, parametric models capture recurring behavioral patterns, and preference- and reward-based methods contribute person-specific evaluative standards.

Several research challenges remain. First, observed decisions are not equivalent to clean preference labels. A decision may reflect external constraints rather than intrinsic preferences, a challenge that is also well known in learning from implicit feedback [38]. Preference- and reward-based methods should therefore represent uncertainty and condition judgments on contextual factors such as audience, role, stakes, deadlines, and available resources. Second, preference evidence is often sparse and uneven across domains. Although personalized and pluralistic reward models improve sample efficiency through shared structure [10, 16, 100], person distillation must ensure that shared information

complements rather than replaces the target person's individuality. Third, preference- and reward-based methods should account for temporal drift by representing how evaluative standards evolve over time instead of collapsing all historical traces into a single static reward function. Finally, preference- and reward-based methods must preserve person specificity. A model that predicts generally good decisions has learned a generic reward rather than a person-specific one. Methods such as matched-peer alternatives, contrastive objectives, and evidence-linked explanations may therefore be necessary to distinguish genuinely person-specific judgment from role-level or population-level preferences.

5.5 Behavioral and Trajectory-Based Distillation

Prompt-based methods represent what a person says, memory-based methods preserve what the person has experienced, parametric and adapter-based methods learn recurring behavioral regularities, and preference- and reward-based methods model how the person evaluates alternatives. Behavioral and trajectory-based distillation focuses instead on *how the person acts over time*. Rather than modeling isolated decisions, it seeks reusable process patterns, such as how a person frames situations, gathers evidence, sequences actions, coordinates with others, revises plans, and responds to uncertainty.

This perspective is closely related to several existing research areas. Agent skill distillation learns reusable procedures from trajectories, tool-use histories, execution feedback, and task outcomes [41, 77, 118]. Process mining discovers workflow structure from event logs [111]. Learning from demonstration and inverse reinforcement learning infer policies or latent objectives from observed behavior [2, 5, 76]. Person distillation shares many of these techniques but changes the learning objective. The goal is not simply to recover a successful workflow or executable skill, but to identify what is distinctive about *this person's* way of acting.

For example, two people may both successfully organize a family trip. One may first collect everyone's constraints, fix the budget, and prepare contingency plans before selecting activities, whereas another may begin with the desired experiences and adjust logistics afterward. From the perspective of task success, both strategies are equally effective. From the perspective of person distillation, however, the ordering of decisions, evidence-seeking habits, coordination strategies, and revision patterns are precisely the signals that distinguish one individual from another.

The first step is therefore to construct *person-centered trajectories*. Rather than treating historical records as isolated events, the system reconstructs coherent episodes consisting of context, evidence, actions, and outcomes. Raw traces are first segmented into episodes, aligned across heterogeneous sources, ordered temporally, and abstracted into higher-level actions such as "collect constraints," "compare alternatives," or "seek confirmation." Existing work on process mining, event-log preprocessing, and event abstraction provides useful techniques for these operations [65, 108, 111–113]. The resulting trajectories should remain linked to their underlying traces so that every abstract action can be traced back to supporting evidence.

Once trajectories have been constructed, the next challenge is to identify *decision points*. Many actions are dictated by the task itself and reveal little about the individual. More informative are situations in which several reasonable actions were available and the observed choice reflects the person's characteristic way of proceeding. Existing decision mining identifies branching points in workflows [92], while inverse reinforcement learning explains observed behavior through latent objectives [2, 76]. Person distillation extends these ideas by asking which choices distinguish the target person from peers performing similar tasks. This often requires reconstructing the alternatives that were available, using matched-peer episodes, workflow templates, or the person's own historical trajectories.

Trajectory information can then be distilled at several levels. *Step-level models* predict the person's next action from the current context, drawing inspiration from learning from demonstration and sequential imitation learning [5, 90].

Sequence-level models learn longer-range behavioral patterns by modeling entire trajectories, similar to Decision Transformer [17]. *Routine-level models* compress frequently occurring action sequences into reusable behavioral routines or options [105]. Regardless of the modeling approach, the objective remains the same: to preserve behaviors that are characteristic of the target person rather than generic task procedures.

Compared with the previous four method families, behavioral and trajectory-based methods capture procedural knowledge that is difficult to express as explicit rules, isolated memories, learned parameters, or preference functions. At the same time, they depend heavily on accurate trajectory reconstruction and often require complementary representations for grounding and interpretation. In practice, behavioral and trajectory-based methods are likely to work together with memory-based retrieval and parametric representations rather than replacing them.

Several research challenges remain. First, action abstraction must balance generalization and fidelity. Actions that are too fine-grained become difficult to learn, whereas actions that are too coarse lose the behavioral distinctions that identify an individual. Hierarchical action vocabularies may provide one solution by preserving links between low-level traces and higher-level behavioral routines. Second, behavioral and trajectory-based methods must generalize beyond sparse and uneven observations without overextending limited evidence. Retrieval-augmented behavioral and trajectory-based methods, uncertainty-aware prediction, and interactive correction mechanisms may help maintain person-specific fidelity. Finally, behavioral and trajectory-based methods must support continual revision. Because behavioral patterns evolve over time, future systems should incorporate time-aware trajectory modeling, change-point detection, and versioned behavioral routines so that new experiences update rather than overwrite a person's evolving process.

5.6 Hybrid and Revisable Person Distillation

The five method families above should be viewed as complementary rather than competing. Prompt-based methods distill a person into an explicit profile, memory-based methods preserve evidence as retrievable memories, parametric and adapter-based methods learn recurring behavioral patterns, preference- and reward-based methods capture evaluative judgment, and behavioral and trajectory-based methods represent procedural behavior. Each preserves different aspects of a person, but each also leaves important gaps. Profiles are transparent but lossy; memories preserve evidence but do not automatically generalize; parametric models learn implicit regularities but are difficult to interpret and revise; preference- and reward-based methods encode judgment but depend on the quality of candidate alternatives; and behavioral and trajectory-based methods capture behavioral processes but require reconstructing incomplete episodes. Faithful person distillation therefore requires combining these complementary capabilities within a unified and continually evolving system.

Current person-distillation systems largely rely on a single representation, such as a Markdown profile or skill file [109, 135]. Adjacent research provides useful building blocks for richer systems. Retrieval-augmented generation combines generation with external evidence [55]; long-term agent memory combines persistent memories with inference-time retrieval [19, 81, 82]; adapter composition and mixture-of-experts methods study how multiple learned modules can be coordinated [29, 84, 99]; and model-editing techniques investigate how learned models can be corrected locally [70, 74, 122]. These methods, however, optimize different objectives. Hybrid person distillation must instead coordinate multiple representations around a single goal: faithfully preserving a person's domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context while remaining evidence-grounded and revisable.

A hybrid person-distillation system should therefore support four core functions. First, it should maintain multiple complementary representations of the same person, including explicit profiles, retrievable memories, learned parametric modules, preference and reward models, and behavioral routines or trajectory models. Second, it should determine which representation, or combination of representations, is appropriate for each query. Third, it should detect and resolve inconsistencies across representations. Finally, it should support continual revision as new traces and feedback become available.

The central methodological challenge is *orchestration*. A hybrid system should not simply concatenate profile text, retrieved memories, adapter outputs, reward scores, and trajectory predictions. Instead, each component should play a well-defined role. One possible strategy is staged composition: retrieve relevant evidence, infer the person's substantive judgment, express that judgment using the person's communication style, and finally verify that the response remains grounded and within the available evidence. This separation of evidence retrieval, reasoning, style realization, and boundary checking makes the overall system more interpretable, debuggable, and revisable.

Hybrid systems must also handle conflicts across representations. A profile may state that a person values careful testing, while recent memories show that they adopted faster release practices after changing roles. Likewise, a judgment model may recommend rejecting a proposal, whereas a communication module prefers a softer presentation. Rather than averaging such disagreements, future systems should explicitly distinguish substantive decisions from their presentation. Judgment modules should determine the person's stance, while communication modules determine how that stance is expressed.

Equally important is *revision*. Person distillation is not a one-time process. People acquire new expertise, change responsibilities, develop new relationships, and revise their own standards. The distilled representation may also contain errors that are corrected by the target person or trusted collaborators. Different representations naturally require different revision mechanisms. Profiles can be edited directly. Memory systems can add, revise, or invalidate individual memories. Parametric models require model editing or additional training [70, 74, 122]. Preference- and reward-based methods require new comparison data or recalibration, while behavioral and trajectory-based methods require additional episodes that reflect updated behavioral patterns. Revision should therefore be component-specific rather than applied uniformly to the entire system.

Revision should also be *facet-specific*. Different aspects of a person evolve at different rates. Technical knowledge may change rapidly; communication style may shift after a promotion; and relationships may evolve as organizational structures change, while core values often remain relatively stable. Updating every representation after each new observation risks destroying stable characteristics, whereas freezing the entire representation causes it to become outdated. Similar to concept drift [30], future systems should therefore support different update strategies for different facets and representation types.

Several research challenges remain. First, systems need principled representation routing that determines which representations should contribute to a given query. Second, they require mechanisms for detecting and explaining inconsistencies across profiles, memories, learned models, preference and reward models, and behavioral or trajectory models. Third, they need local correction methods that revise one component without damaging unrelated person-specific knowledge. Fourth, they should maintain temporally versioned representations that distinguish past behavior from current behavior. Finally, they should preserve revision provenance by recording what changed, why it changed, who provided the correction, and which evidence supported the update.

Taken together, the six method families should be viewed as complementary building blocks rather than independent alternatives. Prompt-based methods provide explicit summaries, memory-based methods preserve evidence, parametric

and adapter-based methods learn implicit regularities, preference- and reward-based methods capture judgment, behavioral and trajectory-based methods represent behavior, and hybrid and revisable methods orchestrate these representations into a coherent, evidence-grounded, and continually evolving computational representation of a real person.

6 An Evaluation Framework for Person Distillation

Evaluation is the final component of person distillation. Unlike conventional AI systems, the objective is not merely to generate plausible or useful outputs, but to faithfully represent a *particular individual*. A system may write like a senior engineer, provide technically sound advice, or successfully complete a task, yet still fail at person distillation if its behavior is generic to the role rather than specific to the target person. The central evaluation question is therefore not *"Is the output good?"* but rather *"Is this what this person would know, say, decide, or do in this situation?"* Answering this question requires evaluating person-specific fidelity rather than correctness or task success alone.

Despite its importance, person distillation currently lacks a principled evaluation framework. Existing systems [23, 109, 116, 135] are primarily presented as repositories, demonstrations, or case studies, without shared benchmarks, standardized evaluation protocols, or a common notion of *person-specific fidelity*. Existing evaluation paradigms address only part of the problem. Role-playing benchmarks evaluate character consistency through dialogue quality, knowledge consistency, utterance style, and emotional expression [110, 134]. Agent benchmarks primarily measure task success [77, 118], while personalized generation benchmarks such as LaMP [94] and PersonalLLM [136] evaluate user adaptation using conventional metrics such as accuracy, F1, MAE/RMSE, and ROUGE. These benchmarks evaluate correctness, task performance, or personalization, but they do not assess whether a system faithfully preserves a person's domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, or relational context. In short, existing benchmarks evaluate whether a system behaves *like someone*, whereas person distillation requires evaluating whether it behaves *like this particular person*.

This section proposes a comprehensive evaluation framework organized around four complementary questions. **Evaluation objectives** define *what* aspects of a person should be evaluated. **Evaluation protocols** define *how* benchmark cases should be constructed. **Evaluation metrics** measure person-specific fidelity, while **diagnostics** assess whether the evaluation itself is trustworthy by examining evidence grounding, boundary awareness, consistency, and robustness. Together, these four components provide a systematic framework for comparing person-distillation systems and measuring fidelity to an individual.

Our framework builds upon established evaluation techniques, including held-out evaluation [39, 50], task success [18, 64, 127], rubric-based judgment [62, 115], authorship comparison [51, 52, 102], hallucination detection [40], selective prediction [31], and inter-annotator agreement [6]. Building upon these foundations, the remainder of this section develops person-distillation-specific evaluation objectives, protocols, metrics, and diagnostics for measuring faithfulness to a particular individual.

6.1 Evaluation Objectives

Section 4 organized distilled person representations around six complementary facets: domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context. These same facets naturally define the evaluation objectives of person distillation. A faithful representation should preserve not only what the target person knows, but also how the person evaluates alternatives, communicates, acts, interacts with others, and where the boundaries of the person's knowledge and preferences lie. Since these facets

capture different aspects of a person, they should be evaluated separately rather than collapsed into a single score. Person distillation therefore requires a *multi-dimensional fidelity profile*.

Knowledge fidelity measures whether the distilled representation faithfully captures the person's knowledge, expertise, and experience while respecting the boundaries of that knowledge. A faithful system should know what the person knows, recognize what the person does not know, and avoid inventing unsupported expertise, experiences, or opinions. Thus, knowledge fidelity evaluates not only factual correctness but also faithfulness to the person's actual experience.

Judgment fidelity measures whether the representation preserves the person's characteristic way of evaluating alternatives and making decisions. Individuals with similar expertise often apply different criteria, prioritize different objectives, and make different tradeoffs. Judgment fidelity therefore asks whether the system reproduces the person's distinctive reasoning rather than merely producing reasonable decisions.

Communication fidelity measures whether the representation expresses ideas in ways that are characteristic of the target person. This includes writing style, tone, organization, level of detail, explanation strategy, and, when available, multimodal behaviors such as speech patterns or visual explanation habits. The objective is to preserve the person's characteristic style of communication across different situations rather than simply matching surface linguistic features [75, 88, 102].

Behavioral fidelity measures whether the representation reproduces the person's recurring patterns of action. Whereas communication fidelity concerns how a person expresses ideas, behavioral fidelity concerns how the person acts. Two people may achieve the same outcome while following different processes. Behavioral fidelity therefore evaluates whether the system preserves characteristic action patterns, evidence-seeking habits, planning strategies, and responses to uncertainty.

Values and personality fidelity measures whether the representation consistently reflects the person's values, preferences, priorities, risk tolerance, interpersonal disposition, and personal boundaries. These characteristics often influence decisions even when factual knowledge is identical, making them an essential component of person-specific behavior.

Relational fidelity measures whether the representation adapts appropriately to different social relationships and organizational contexts. People naturally communicate and behave differently with close collaborators, junior colleagues, senior leadership, customers, or external partners. Relational fidelity therefore evaluates whether these relationship-dependent patterns are preserved rather than treating behavior as independent of social context.

Together, these six objectives define a comprehensive notion of *person fidelity*. Unlike conventional evaluation, which typically focuses on task success, personalization, or role consistency, person distillation evaluates whether a system faithfully represents a *particular individual* across multiple complementary dimensions. The following sections translate these objectives into practical evaluation protocols, quantitative metrics, and diagnostic analyses.

6.2 Benchmark Construction

Section 6.1 defined *what* aspects of a distilled person should be evaluated. This section addresses the complementary question of *how* those evaluations should be designed. Unlike many machine learning problems, person distillation cannot be evaluated using a single benchmark or experimental protocol. Different evaluation protocols probe different aspects of person fidelity, making benchmark construction an experimental methodology rather than a single dataset or benchmark.

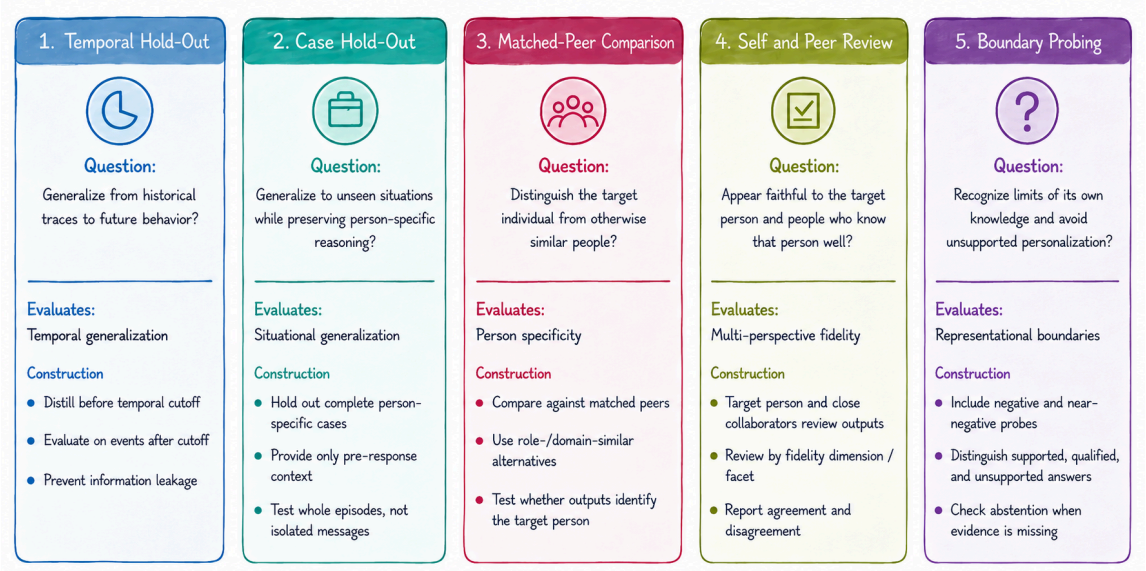

Fig. 6. Benchmark construction protocols for person distillation. The five protocols evaluate complementary aspects of person fidelity: temporal generalization, situational generalization, person specificity, multi-perspective fidelity, and representational boundaries.

We organize benchmark construction around five complementary protocols (Figure 6). **Temporal hold-out** evaluates whether a distilled representation generalizes from historical traces to future behavior. **Case hold-out** evaluates generalization to previously unseen situations while preserving person-specific reasoning. **Matched-peer comparison** evaluates whether the representation distinguishes the target individual from similar people. **Self and peer review** evaluates whether the representation appears faithful to both the target person and knowledgeable collaborators. Finally, **boundary probing** evaluates whether the representation recognizes the limits of its knowledge and avoids unsupported personalization. These protocols are complementary rather than competitive: each reveals different strengths and failure modes, and together they provide a more comprehensive evaluation of person-specific fidelity than any single benchmark alone.

6.2.1 Temporal Hold-Out. **Temporal hold-out** evaluates whether a distilled representation can generalize from historical traces to the future behavior of a particular person. The representation is constructed using only traces collected before a temporal cutoff and is evaluated on events that occur afterward. This follows the standard principle of time-aware evaluation that training data should contain only information available before the prediction target [39]. Similar chronological protocols are widely used in sequential recommendation, where each user's most recent interactions are held out for evaluation [36, 45]. Person distillation extends this idea from predicting future user behavior to predicting the future behavior of a *particular individual*.

A temporal benchmark partitions a person's traces into a *distillation window* before a cutoff time t_c and an *evaluation window* afterward. The distillation window contains all evidence available before t_c, such as code reviews, design documents, incident reports, chat messages, and meeting notes. The evaluation window contains person-specific events that occur later. To avoid *information leakage* [46], the system should receive only the information that was available to the target person at the time of each evaluation event. For example, when evaluating a code review, the model should

receive the pull request before the review was written; when evaluating an incident response, it should receive the system state before any actions were taken. Later discussions, outcomes, and retrospective analyses should be excluded.

Temporal hold-out naturally evaluates multiple aspects of person fidelity within a single benchmark. A held-out code review can assess not only whether the system reaches the same decision, but also whether it identifies similar concerns, requests comparable evidence, and communicates in a manner consistent with the target person. Likewise, a held-out incident can evaluate both the final decision and the sequence of actions used to reach it. Compared with conventional prediction tasks that focus on a single output label, temporal hold-out provides a richer evaluation of knowledge, judgment, communication, and behavior.

A rigorous temporal benchmark should clearly report its construction, including the distillation window, the validation window, the evaluation window, the temporal gap between training and evaluation, and whether the split is performed globally or separately for each individual. Beyond a single train–test split, a stronger protocol is *rolling temporal evaluation*, where the representation is repeatedly distilled from traces before successive cutoff times and evaluated on subsequent events. Rolling evaluation measures not only predictive accuracy but also the temporal stability of the distilled representation.

The principal limitation of temporal hold-out is *temporal drift*. People continuously acquire new knowledge, assume new responsibilities, adopt new tools, and revise their decision-making habits. Consequently, degraded performance may reflect deficiencies in the distilled representation, genuine evolution of the person, or both. This challenge is closely related to *concept drift* in machine learning [30]. Temporal benchmarks should therefore document major changes in the target person's role, organization, or working context during the evaluation period.

More broadly, temporal hold-out raises an important research question: should a distilled representation preserve the *historical* person or continually evolve toward the *current* person? Short temporal gaps primarily evaluate near-term prediction, whereas longer gaps increasingly measure whether the representation remains faithful as the individual changes. Understanding this tradeoff between *person fidelity* and *person evolution* is likely to become a central research problem for continually updated person-distillation systems.

6.2.2 Case Hold-Out. Whereas temporal hold-out evaluates generalization across time, **case hold-out** evaluates generalization across situations. The goal is to determine whether a distilled representation has learned the target person's characteristic way of reasoning rather than memorizing previously observed episodes. Instead of holding out all traces after a temporal cutoff, the evaluator removes one or more complete cases from the person's history and reserves them exclusively for testing. This follows the general principle of held-out evaluation and cross-validation [50], but extends it from independent examples to person-specific episodes.

A *case* is a coherent unit of activity, such as a code review, design discussion, incident response, feature-prioritization meeting, mentoring interaction, or conflict-resolution process. Unlike evaluating individual messages or actions, treating the entire episode as the evaluation unit preserves the dependencies among context, reasoning, communication, decisions, and outcomes.

During evaluation, the system receives only the information that was available before the target person's response within the held-out case. For example, in a code review task, the system receives the pull request and any preceding discussion, but not the review itself. In a design discussion, it receives the proposal and supporting materials, but not the person's subsequent comments or decision. The evaluation target may include the person's response, final decision, action sequence, or recorded rationale. The objective is therefore not to reproduce historical text, but to determine

whether the distilled representation behaves as the target person would in a previously unseen instance of a familiar problem.

Compared with temporal hold-out, case hold-out is less sensitive to temporal drift because the held-out episode may originate from any point in the person's history. Instead, it focuses more directly on *person-specific reasoning*. For example, two senior engineers may receive the same deployment proposal yet consistently differ in whether they approve it immediately, request additional monitoring, postpone deployment, or require a rollback strategy. Case hold-out evaluates whether the representation reproduces these characteristic tendencies rather than merely producing a technically reasonable recommendation.

The principal challenge is *case-level leakage*. A held-out case is often connected to follow-up emails, meeting summaries, retrospective reports, documentation updates, or later conversations that explicitly or implicitly reveal the person's decisions. If these artifacts remain in the distillation corpus, the system may recover the correct answer through leaked evidence rather than person-specific reasoning [46]. Consequently, benchmark construction should clearly define what constitutes a case, identify all related artifacts, and exclude information that reveals the outcome of the held-out episode.

More broadly, case hold-out shifts the evaluation objective from memorization to reasoning. Rather than asking whether a system can reproduce previously observed outputs, it asks whether the distilled representation has learned how the target person analyzes unfamiliar situations, weighs alternatives, and reaches decisions. Together with temporal hold-out, case hold-out provides a complementary view of generalization: the former evaluates fidelity across time, whereas the latter evaluates fidelity across situations.

6.2.3 Matched-Peer Comparison. Temporal hold-out evaluates generalization across time, and case hold-out evaluates generalization across situations. **Matched-peer comparison** addresses a different question: *does the distilled representation capture what makes the target person unique?* Even if a system accurately predicts future behavior and generalizes to unseen cases, it may still behave like a generic individual with similar expertise, role, or experience. Matched-peer comparison therefore evaluates **person specificity** by determining whether the representation can be distinguished from representations of *similar people*. This idea is closely related to authorship attribution and verification, where an unknown document is compared against candidate authors or same-author/different-author examples [51, 52, 102].

The benchmark is constructed by comparing the distilled representation against carefully matched alternatives. Given a held-out scenario, such as a pull request, design proposal, incident response, or mentoring situation, the evaluator prepares multiple candidate responses. These may include the target person's actual response, the response generated by the distilled representation, responses from matched peers, and outputs from generic baselines such as role-only prompting or retrieval-only systems. Human reviewers or automatic evaluators then identify or rank the responses according to how well they match the target person. Comparative annotation methods such as best–worst scaling are particularly suitable because they reduce ambiguity compared with absolute rating scales [49].

Unlike the previous two protocols, matched-peer comparison explicitly measures *distinguishability*. Temporal and case hold-out ask whether a representation predicts the target person's future behavior or reasoning. Matched-peer comparison asks whether those behaviors remain distinguishable from plausible alternatives. This distinction is important because people with similar expertise often differ systematically in judgment, communication style, risk tolerance, mentoring approach, or leadership philosophy. A faithful person representation should therefore preserve the characteristics that differentiate one individual from others who appear superficially similar.

An important extension is **cohort-based comparison**. Instead of comparing only a few matched peers, the evaluator constructs distilled representations for a larger population and compares them collectively. This idea is related to the intuition behind k-anonymity [106]. If a representation cannot be distinguished from a sufficiently large group of similar people, it may provide stronger privacy protection, but it also provides weaker evidence that person-specific characteristics have been preserved. The goal is therefore to understand the balance between *cohort coherence* and *individual distinctiveness*.

A practical implementation exposes every distilled representation to the same collection of evaluation probes spanning all dimensions of person fidelity. Responses may then be represented as embeddings or rubric-based evaluation vectors and compared across individuals. Pairwise similarities, clustering methods, and visualization techniques can reveal whether meaningful population structure emerges. Standard clustering diagnostics, such as silhouette analysis [91], provide one way to quantify whether the resulting representations exhibit both *cohort coherence* and *individual separation*.

The effectiveness of matched-peer comparison depends critically on benchmark design. Weak or obviously different baselines artificially inflate person specificity. Matched peers should therefore be selected to be genuinely confusable with the target person by sharing similar roles, experience, technical domains, organizational contexts, or communication environments. Evaluation reports should clearly document peer-selection criteria, the number of candidate responses, whether evaluators were blinded to response sources, and whether person specificity was measured globally, facet by facet, or through cohort-level analyses.

Together with temporal and case hold-out, matched-peer comparison completes the evaluation of generalization by asking a complementary question. Rather than evaluating whether a representation behaves correctly or consistently, it evaluates whether the behavior is *uniquely identifiable as that of the particular person*. Person specificity is therefore not simply another evaluation metric, but a defining property of person distillation itself.

6.2.4 Self and Peer Review. Temporal hold-out, case hold-out, and matched-peer comparison evaluate person fidelity by comparing a distilled representation against historical evidence and observable behavior. **Self and peer review** provides a complementary perspective by asking whether the representation is recognizable to the target person and to people who know that person well. This protocol is particularly valuable for evaluating judgment, communication style, behavioral tendencies, values, personality, and relational behavior, where historical traces often provide only partial evidence.

The key assumption is not that human judgments provide perfect ground truth, but that they provide complementary perspectives. Human judgments are inherently subjective and may disagree [130]. The target person may misremember past decisions, describe themselves in socially desirable ways, or overlook recurring patterns that are evident in their historical traces. Likewise, different collaborators observe different aspects of the same individual depending on their roles, relationships, and shared experiences. This observation is consistent with research showing that self-reports and observer reports capture complementary rather than identical aspects of personality [21, 114]. Consequently, disagreement among reviewers should be interpreted as evidence about different perspectives rather than simply as annotation noise.

A practical benchmark can again be constructed using held-out cases. For each case, the target person and one or more knowledgeable collaborators evaluate whether the distilled response faithfully reflects the person's knowledge, judgment, communication style, behavioral patterns, values, and relational behavior. Structured rubrics should encourage reviewers to assess each fidelity dimension independently rather than assigning a single overall score. Reviewers should also be allowed to indicate *"not enough evidence,"* recognizing that a response may appear plausible while remaining

unsupported by the available traces. When expert evaluation becomes expensive, active learning can prioritize the most informative cases, focusing human effort on uncertain, highly person-specific, or potentially controversial examples [97].

Compared with the previous protocols, self and peer review evaluates aspects of person fidelity that are difficult to measure automatically. Historical traces reveal what a person did, but they do not always explain how those actions were perceived by the person or by others. For example, a code-review history may reveal technical decisions, while close collaborators may better recognize whether the person's mentoring style, interpersonal behavior, or decision-making philosophy has been faithfully reproduced. Human evaluation therefore complements evidence-based evaluation by assessing person-specific dimensions that are difficult to operationalize through automatic metrics.

Because human evaluation is itself part of the benchmark, it should also be evaluated systematically. Benchmark reports should document who participated, how reviewers were selected, what relationship they had with the target person, which fidelity dimensions they evaluated, whether they had access to supporting evidence, and how disagreements were handled. Self-ratings and peer ratings should be reported separately before aggregation because they capture different perspectives on the same individual. Such transparency improves reproducibility and enables future work to study when different observers agree, when they disagree, and which aspects of person fidelity are most sensitive to perspective.

Together with temporal hold-out, case hold-out, and matched-peer comparison, self and peer review provides an essential human-centered perspective on evaluation. Rather than asking only whether a representation predicts behavior or distinguishes one individual from another, it asks whether the representation is recognized as a faithful representation of the target person. In this sense, disagreement is not merely an evaluation challenge, but an informative signal about the inherently multi-perspective nature of human identity.

6.2.5 Boundary Probing. The previous benchmark protocols evaluate whether a distilled representation faithfully reproduces what a person would know, say, decide, or do. **Boundary probing** evaluates the complementary question: *does the representation also recognize what the person does not know?* Rather than rewarding a system for answering every question, boundary probing evaluates whether it appropriately abstains, qualifies its conclusions, or explicitly acknowledges uncertainty when the available evidence is insufficient. This perspective is closely related to hallucination detection and selective prediction, where systems are expected to avoid unsupported claims when reliable evidence is unavailable [31, 40].

A practical benchmark consists of both *negative* and *near-negative* probes. Negative probes concern information that is absent from the person's traces, such as projects they never worked on, opinions they never expressed, decisions made by other people, or events that occurred after the distillation window. Near-negative probes are more challenging because they concern topics adjacent to the person's expertise while remaining unsupported by evidence. For example, a backend engineer may have extensive experience with payment reliability, but this does not justify inferring their opinion about every payment-related design decision. Likewise, a manager who frequently mentors junior engineers should not automatically be assumed to have a particular relationship with every employee. These probes evaluate whether the representation distinguishes reasonable generalization from unsupported personalization.

Boundary probing complements the previous evaluation protocols by explicitly measuring the limits of person fidelity. Temporal hold-out evaluates generalization across time, case hold-out evaluates generalization across situations, matched-peer comparison evaluates person specificity, and self and peer review evaluates perceived faithfulness. Boundary probing instead evaluates whether every person-specific claim remains supported by evidence. This distinction

is particularly important because person distillation must generalize beyond observed traces without fabricating unsupported beliefs, preferences, relationships, or experiences [131].

The correct behavior is not always refusal. In many situations, the most faithful response is a *qualified answer*. For example, a distilled engineer might respond, "The person consistently emphasized observability in similar production incidents, but there is no evidence regarding this particular service." Such responses appropriately generalize from historical evidence while making the limits of that evidence explicit. Benchmark construction should therefore distinguish among supported answers, qualified answers, unsupported but plausible answers, and clear hallucinations. Whenever possible, person-specific claims should also be linked to the evidence from which they were inferred.

Boundary probing highlights a defining characteristic of person distillation: ignorance is itself part of a faithful representation. Every individual has limits on their knowledge, experience, memories, and relationships. A representation that confidently answers every question may appear capable while fabricating unsupported personal information. Future benchmark construction should therefore report the composition of boundary probes, the proportion of probes that require abstention or qualification, and the frequency of unsupported person-specific claims. Modeling the boundaries of a person's knowledge is as important as modeling the knowledge itself.

Taken together, these five benchmark construction protocols provide complementary views of person fidelity. Temporal hold-out evaluates generalization across time, case hold-out evaluates generalization across situations, matched-peer comparison evaluates person specificity, self and peer review evaluates human-perceived fidelity, and boundary probing evaluates whether the representation respects the limits of available evidence. No single protocol is sufficient on its own. Person distillation should therefore be viewed as a **multi-protocol evaluation problem**, where different protocols reveal complementary strengths and failure modes [130].

An important direction beyond these benchmark protocols is **cross-source and cross-representation evaluation**. Rather than evaluating only a single distilled representation, future work should compare representations constructed from different source traces and representation formats. For example, one may compare representations distilled from code reviews, chat histories, documents, or meeting recordings, as well as document-based representations, parametric model representations, preference and reward model representations, and memory and knowledge system representations. Agreement across independently constructed representations may indicate stable person-specific characteristics rather than artifacts of a particular data source or representation. Conversely, disagreement is not necessarily a failure. Different sources naturally capture different facets of a person: chat histories may better reveal communication style and relational context, whereas code reviews and design documents may better capture technical judgment and decision criteria. Future evaluation should therefore examine not only whether different representations agree, but also *which aspects of person fidelity remain stable across sources and which are inherently source-dependent*. Such analyses will deepen our understanding of both person distillation and the nature of computational person representations.

6.3 Evaluation Metrics

Section 6.2 defined how evaluation benchmarks should be constructed. This section addresses the complementary question of *how* performance on those benchmarks should be measured. Existing evaluation paradigms provide useful measurement foundations. Role-playing benchmarks measure character consistency, communication style, and conversational believability [110, 120]; agent benchmarks measure task success, reward, progress, and action validity [64, 127]; and code-generation benchmarks execute generated programs against test cases [18]. Person distillation

builds upon these ideas while asking a different question: *does the output faithfully represent what this particular person would know, decide, say, or do?*

We organize the metrics into three levels. **Metric foundations** define the basic measurement primitives used to compare system outputs with the target person's observed behavior. **Person-fidelity metrics** quantify how faithfully the distilled representation preserves the target person's domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context. **Person specificity metrics** measure whether the representation captures characteristics that distinguish the target individual from strong generic baselines and matched peers. Together, these metrics quantify person fidelity, while the diagnostic analyses in Section 6.4 assess whether the resulting measurements are trustworthy.

Throughout this section, let p denote the target person, R_p the distilled representation, x_i a held-out evaluation case, y_i^p the person's observed response, decision, or action, and $\hat{y}_i = R_p(x_i)$ the system output. Different metrics are computed over different evaluation sets, denoted by $\mathcal{E}$ with appropriate subscripts.

Label matching. The simplest measurement primitive compares the system output with the person's observed decision when the evaluation target is a discrete label, such as approve/request changes, escalate/wait, or accept/reject. When outputs are free-form, the corresponding labels may first be extracted by human annotators or LLM-based classifiers before comparison.

Execution or state checking. Some outputs can be executed directly in an environment. Examples include running generated programs against unit tests or executing action sequences in a simulated environment. For an execution-based evaluation set $\mathcal{E}_{\text{exec}}$, the task success rate is

$$\text{SR}(p) = \frac{1}{|\mathcal{E}_{\text{exec}}|} \sum_{i \in \mathcal{E}_{\text{exec}}} \mathbf{1}\{\text{Check}(x_i, \hat{y}_i) = 1\},$$

where Check denotes a task-specific verifier, such as a unit-test suite, an environment reward function, a final-state checker, or a subgoal-completion checker. Although execution success is an important measure of task performance, it does not by itself establish person fidelity.

Reference-based similarity. When the held-out case contains the person's observed response y_i^p, the generated output $\hat{y}_i$ can be compared using text overlap, embedding similarity, stylometric features, action-sequence similarity, or edit distance [130]. Because multiple faithful responses may exist, these similarity scores should be interpreted as evidence of consistency rather than exact correctness.

Rubric-based judgment. Many aspects of person fidelity cannot be reduced to labels, execution success, or textual similarity. In these cases, outputs are evaluated using explicit rubrics by human reviewers, peers, LLM-as-judge systems, or trained reward models [62, 115]. Rather than assigning a single overall score, the rubric should specify the particular aspect being evaluated, such as judgment, evidence use, communication style, relational adaptation, or boundary awareness. This primitive provides a common foundation for many of the person-fidelity metrics introduced below.

Outcome metrics. The most direct evaluation asks whether the distilled representation reaches the same observable outcomes as the target person. The appropriate metric depends on the task. When the outcome is a discrete decision, such as approve/request changes, escalate/wait, or accept/reject, *decision alignment* measures whether the system reproduces the person's observed decision. Let $\mathcal{E}_{\text{D}}$ denote the evaluation set of held-out decision cases. If $d(y_i^p)$ and $d(\hat{y}_i)$ denote the decisions made by the target person and the distilled representation, respectively, then

$$\mathrm{DA}(p) = \frac{1}{|\mathcal{E}_{\mathrm{D}}|} \sum_{i \in \mathcal{E}_{\mathrm{D}}} \mathbf{1}\{d(\hat{y}_i) = d(y_i^p)\}.$$

Decision alignment differs from ordinary classification accuracy because the objective is not to recover the objectively correct decision, but the decision that the *target person* would make.

Some tasks additionally admit an external notion of success. For example, generated code can be executed against unit tests, and interactive agents can be evaluated by whether they complete a task successfully. Such metrics remain valuable because they quantify functional correctness. A common example is pass@k [18], which estimates the probability that at least one of k sampled programs passes all test cases:

$$\text{pass@}k = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}},$$

where n is the number of sampled programs and c is the number of correct ones. However, external task success should be interpreted as complementary rather than sufficient. A system may complete the task successfully while following reasoning, design choices, or communication patterns that differ substantially from those of the target person. Consequently, task-success metrics should always be reported together with person-centered metrics.

Judgment alignment. Many person-specific decisions cannot be evaluated by exact agreement alone. Two responses may reach the same conclusion for different reasons, while two different conclusions may nevertheless reflect similar reasoning. Judgment alignment therefore evaluates *how* the decision was reached rather than only *what* decision was made. Building upon rubric-based evaluation [62, 115], we propose evaluating four complementary aspects of judgment: whether the output identifies the same primary concern, uses similar evidence, prioritizes considerations in a similar order, and reaches a similar final stance or action.

Let $\mathcal{E}_{\mathrm{J}}$ denote the evaluation set for judgment. For each case i, let $q_{\text{concern},i}$, $q_{\text{evidence},i}$, $q_{\text{priority},i}$, and $q_{\text{stance},i}$ denote rubric scores in $[0, 1]$. A judgment alignment score is

$$\mathrm{JA}_i = w_1 q_{\text{concern},i} + w_2 q_{\text{evidence},i} + w_3 q_{\text{priority},i} + w_4 q_{\text{stance},i}, \qquad \sum_{j=1}^{4} w_j = 1,$$

and the overall judgment alignment is

$$\mathrm{JA}(p) = \frac{1}{|\mathcal{E}_{\mathrm{J}}|} \sum_{i \in \mathcal{E}_{\mathrm{J}}} \mathrm{JA}_i.$$

The weighting scheme should be reported explicitly because different applications emphasize different aspects of judgment. For example, safety-critical evaluations may place greater weight on the primary concern and final decision, whereas evaluations emphasizing interpretability may assign higher weight to evidence use and priority ordering. Compared with decision alignment, judgment alignment provides a richer characterization of person-specific reasoning.

Behavioral trajectory alignment. Many person-distillation tasks require evaluating not a single decision, but an entire sequence of actions. Existing agent benchmarks typically measure task success, accumulated reward, or progress toward a goal [64, 127]. Person distillation introduces an additional requirement: when multiple successful action sequences exist, the distilled representation should follow a trajectory that resembles the target person's characteristic way of working.

Let $\mathcal{E}_{\mathrm{T}}$ denote the set of held-out trajectory cases. For each case, let $a^{p}_{i,1:T_i}$ be the person's observed action sequence and $\hat{a}_{i,1:\hat{T}_i}$ the system-generated sequence. We measure similarity using both subgoal completion and action-order similarity:

$$\mathrm{TA}_i = \alpha\,\mathrm{SubgoalF1}(a^{p}_{i,1:T_i}, \hat{a}_{i,1:\hat{T}_i}) + (1-\alpha)\left(1 - \mathrm{NED}(a^{p}_{i,1:T_i}, \hat{a}_{i,1:\hat{T}_i})\right),$$

where SubgoalF1 measures overlap between completed subgoals, NED is the normalized edit distance between the two action sequences, and $\alpha \in [0,1]$ controls the relative importance of the two components. The average trajectory alignment is

$$\mathrm{TA}(p) = \frac{1}{|\mathcal{E}_{\mathrm{T}}|}\sum_{i\in\mathcal{E}_{\mathrm{T}}} \mathrm{TA}_i.$$

Behavioral trajectory alignment complements both task success and decision alignment. Two systems may successfully complete the same task and even make the same final decision, yet differ substantially in how they gather evidence, interact with collaborators, revise plans, or respond to uncertainty. Trajectory alignment therefore measures whether the distilled representation preserves the target person's characteristic process rather than merely reproducing the final outcome.

Facet-level fidelity metrics. Section 6.1 defined person fidelity as a multi-dimensional concept consisting of domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context. Evaluation should therefore report fidelity for each facet rather than collapsing all results into a single score. Let $\mathcal{F}$ denote the set of evaluated facets, and let $\mathcal{P}_f$ be the probe set for facet $f \in \mathcal{F}$. Depending on the facet, the measurement primitive may be label matching, execution or state checking, reference-based similarity, or rubric-based judgment. Let $m_f(x_i, \hat{y}_i, y^{p}_i) \in [0,1]$ denote the corresponding measurement function. The fidelity score for facet f is

$$S_f(p) = \frac{1}{|\mathcal{P}_f|}\sum_{i\in\mathcal{P}_f} m_f(x_i, \hat{y}_i, y^{p}_i).$$

Rather than reporting only an overall average, person distillation should report the complete fidelity profile

$$\mathrm{S}(p) = \big(S_f(p)\big)_{f\in\mathcal{F}},$$

which reveals the strengths and weaknesses of the distilled representation across different aspects of a person. For example, a representation may achieve high knowledge fidelity while exhibiting weaker judgment or relational fidelity. Such a profile is substantially more informative than a single aggregate score because it identifies *what* has been successfully distilled and *what* remains difficult to capture.

Facet-level evaluation can also be organized by depth. Level 1 probes evaluate broad domain knowledge, Level 2 probes evaluate specialized expertise, Level 3 probes evaluate organization-specific knowledge, Level 4 probes evaluate team-specific practices, and Level 5 probes evaluate person-specific judgment and behavior. Reporting fidelity as a depth curve helps distinguish systems that capture only general domain knowledge from those that preserve increasingly individual characteristics.

Communication and relational similarity. Communication fidelity evaluates whether the system expresses ideas in ways that resemble the target person. Existing work in authorship attribution and stylometry provides useful measurement techniques for comparing writing style, lexical choice, sentence structure, and other stylistic

characteristics [75, 102]. More recent approaches compare learned authorship or text embeddings to capture broader semantic and stylistic patterns [130]. These methods provide useful measurement primitives, but communication fidelity extends beyond writing style to include explanation strategy, level of detail, tone, and adaptation to different audiences.

Let $\mathcal{E}_{\mathrm{C}}$ denote the communication evaluation set. Let $\phi(\cdot)$ denote a communication representation, such as a stylometric feature vector or a learned embedding, and let sim denote an appropriate similarity function. The raw communication similarity for case i is

$$\mathrm{CS}_i = \mathrm{sim}\left(\phi(\hat{y}_i), \phi(y_i^p)\right),$$

and the average communication similarity is

$$\mathrm{CS}(p) = \frac{1}{|\mathcal{E}_{\mathrm{C}}|} \sum_{i \in \mathcal{E}_{\mathrm{C}}} \mathrm{CS}_i.$$

Raw similarity, however, is often insufficient because people in similar professional roles naturally share vocabulary, conventions, and writing styles. A stronger metric therefore compares the generated response with matched-peer outputs. Let $\mathcal{B}_{\mathrm{peer}}(i)$ denote the index set of matched-peer responses for case i. For each peer response,

$$s_{i,b}^{\mathrm{peer}} = \mathrm{sim}\left(\phi(y_{i,b}^{\mathrm{peer}}), \phi(y_i^p)\right), \qquad b \in \mathcal{B}_{\mathrm{peer}}(i),$$

and let $\mu_{\mathrm{peer}}(i)$ and $\sigma_{\mathrm{peer}}(i)$ denote their mean and standard deviation. The normalized communication similarity is

$$\mathrm{NCS}_i = \frac{\mathrm{CS}_i - \mu_{\mathrm{peer}}(i)}{\sigma_{\mathrm{peer}}(i) + \epsilon},$$

where $\epsilon > 0$ avoids division by zero. Averaging over all communication cases gives

$$\mathrm{NCS}(p) = \frac{1}{|\mathcal{E}_{\mathrm{C}}|} \sum_{i \in \mathcal{E}_{\mathrm{C}}} \mathrm{NCS}_i.$$

Positive values indicate that the generated response is more similar to the target person than typical matched-peer responses, whereas values near zero indicate that the response is no more distinctive than those of similar individuals. The same idea naturally extends to *relational fidelity* by comparing responses within different relationship contexts, such as teammates, managers, junior colleagues, customers, or external collaborators.

Taken together, these person-fidelity metrics evaluate complementary aspects of a distilled representation. Outcome metrics measure whether the system reaches the same decisions as the target person, judgment alignment evaluates whether it reasons in similar ways, trajectory alignment measures whether it follows similar behavioral processes, facet-level metrics reveal which aspects of the person are faithfully preserved, and communication and relational metrics assess how the person's characteristic style is expressed across different social contexts. The next section complements these fidelity measures with diagnostics that assess whether the evaluation is trustworthy.

6.4 Evaluation Diagnostics

The metrics in Section 6.3 quantify how faithfully a distilled representation matches the target person. High metric values, however, do not necessarily imply that the evaluation is trustworthy. A representation may achieve strong decision alignment while relying on unsupported evidence, overgeneralizing beyond the available traces, or exploiting

artifacts of a particular benchmark. Evaluation therefore requires not only fidelity metrics but also *diagnostics* that explain why a representation succeeds or fails and whether the reported measurements can be trusted.

We organize evaluation diagnostics into four complementary categories. **Evidence-grounding diagnostics** examine whether person-specific claims are supported by the available traces. **Boundary diagnostics** evaluate whether the representation appropriately distinguishes supported conclusions from unsupported speculation. **Consistency diagnostics** examine whether different evidence sources and representation formats produce compatible conclusions. Finally, **evaluation-reliability diagnostics** assess the stability and reliability of the evaluation process itself.

Evidence-grounding diagnostics. Person distillation should produce claims that remain traceable to the underlying evidence rather than merely sounding plausible. Let $\mathcal{E}_{\text{claim}}(p)$ denote the evaluation outputs analyzed for claim-level grounding for target person p, and let $C(\hat{y}_i)$ denote the set of person-specific claims contained in output $\hat{y}_i$. Each claim γ is labeled as directly supported, weakly supported, or unsupported by the available traces:

$$e(\gamma; D_p) = \begin{cases} 1, & \text{directly supported,} \\ 0.5, & \text{weakly supported,} \\ 0, & \text{unsupported.} \end{cases}$$

The evidence support rate is

$$\text{ESR}(p) = \frac{\sum_{i \in \mathcal{E}_{\text{claim}}(p)} \sum_{\gamma \in C(\hat{y}_i)} e(\gamma; D_p)}{\sum_{i \in \mathcal{E}_{\text{claim}}(p)} |C(\hat{y}_i)|},$$

while the unsupported claim rate is

$$\text{UCR}_{\text{claim}}(p) = \frac{\sum_{i \in \mathcal{E}_{\text{claim}}(p)} \sum_{\gamma \in C(\hat{y}_i)} \mathbf{1}\{e(\gamma; D_p) = 0\}}{\sum_{i \in \mathcal{E}_{\text{claim}}(p)} |C(\hat{y}_i)|}.$$

Together, these diagnostics measure whether person-specific conclusions remain grounded in supporting evidence rather than relying on unsupported personalization.

Boundary diagnostics. A faithful representation should know not only what the target person knows but also what the available evidence does not justify. Building on hallucination detection and selective prediction [31, 40], we distinguish supported probes $Q^+(p)$ from unsupported probes $Q^-(p)$. Balanced boundary accuracy is

$$\text{BBA}(p) = \frac{1}{2}\left(\frac{1}{|Q^+(p)|} \sum_{i \in Q^+(p)} \mathbf{1}\{\text{supported answer}\} + \frac{1}{|Q^-(p)|} \sum_{i \in Q^-(p)} \mathbf{1}\{\text{qualified answer or abstention}\} \right),$$

which rewards systems that both answer supported questions and refrain from unsupported personalization. This diagnostic complements the benchmark protocol of Section 6.2 by measuring how well the representation respects the limits of its knowledge.

Consistency diagnostics. Faithful person representations should remain reasonably consistent across different evidence sources and representation formats. Let $R_p^{(s)}$ and $R_p^{(t)}$ denote two representations of the same person constructed from different sources or formats. On a shared probe set $\mathcal{P}_{\text{shared}}(p)$,

$$\text{CSC}_p(s,t) = \frac{1}{|\mathcal{P}_{\text{shared}}(p)|} \sum_{i \in \mathcal{P}_{\text{shared}}(p)} \text{sim}_g\left(g(R_p^{(s)}, x_i), g(R_p^{(t)}, x_i) \right),$$

where the similarity function is chosen according to the output type. High consistency suggests stable person-specific characteristics, whereas systematic disagreement may indicate that different sources capture different facets of the individual. Consequently, consistency should be analyzed both globally and separately for each person-fidelity facet.

Evaluation-reliability diagnostics. Many person-fidelity metrics rely on human reviewers, peers, LLM-as-judge systems, or learned reward models. The evaluation itself should therefore be assessed for reliability. LLM judges may exhibit position bias, verbosity bias, or self-preference bias [121], making it important to compare their judgments against human evaluations. Human studies should report inter-annotator agreement [6], while LLM judges and learned evaluators should report correlation with human or peer judgments [62]. For self and peer review, self-ratings and peer ratings should be reported separately before aggregation because they provide complementary perspectives on the same individual.

Taken together, these diagnostics complement the fidelity metrics introduced in Section 6.3. Metrics quantify how faithfully a representation captures a particular person, whereas diagnostics explain whether those measurements are grounded, well calibrated, internally consistent, and reliable. Reporting both provides a substantially more complete evaluation than either metrics or diagnostics alone.

7 Conclusion and Future Directions

This paper establishes the conceptual foundations of **person distillation** and, to our knowledge, presents the first comprehensive survey of this emerging research area. Rather than viewing role-playing, personalization, memory systems, skill learning, and preference modeling as independent research directions, we unify them under a common problem: *transforming heterogeneous traces of a particular real person into faithful, evidence-grounded, and revisable computational representations*. By organizing the field as an end-to-end lifecycle spanning **source evidence**, **distilled person representations**, **distillation methods**, and **evaluation**, we provide a common vocabulary, a coherent conceptual framework, and a reference architecture for understanding how these components jointly determine the fidelity of a distilled person.

More broadly, this work argues that **person distillation is fundamentally different from adapting models to users or imitating personas**. The central challenge is not simply to reproduce how a person writes or acts, but to faithfully preserve what distinguishes that individual: their domain knowledge and expertise, judgment and decision criteria, communication style, behavioral patterns, values and personality, and relational context, while remaining grounded in available evidence and explicitly recognizing the limits of what can be inferred. This perspective shifts attention from isolated algorithms toward the complete lifecycle of representing a person, emphasizing that *source evidence, distilled person representations, distillation methods, and evaluation must be designed as a coherent system rather than optimized independently*.

We hope this framework provides a foundation for future research in person distillation. By identifying the fundamental design principles, open problems, and connections across previously fragmented research areas, it offers a roadmap for developing AI systems that can faithfully, transparently, and responsibly represent particular individuals. As AI assistants become increasingly personalized, collaborative, and long-lived [59], we believe **person distillation will become a foundational capability for the next generation of human-centered intelligent systems**.

The preceding sections identify open challenges within each stage of the person-distillation lifecycle. Looking ahead, however, we believe that the most significant research opportunities lie *beyond the lifecycle itself*. As person representations become increasingly persistent, reusable, and capable, they will no longer function merely as machine learning artifacts. Instead, they will become **computational entities** that interact with people, organizations, and

other AI systems over extended periods of time. This transition fundamentally expands the scope of person distillation. The central questions are no longer limited to *how to construct* faithful person representations, but extend to *how such representations should evolve, collaborate, be governed, and ultimately participate in future AI ecosystems*. We believe the following directions will define the next stage of research.

Person representations should evolve from static artifacts into living digital counterparts. Current systems largely construct static representations from historical traces. Future systems should instead model people as continually evolving entities [72]. Knowledge accumulates, judgment matures, communication styles adapt, relationships change, and professional responsibilities evolve throughout a person's lifetime [89, 95]. Consequently, person distillation should become a *lifelong process* rather than a one-time construction task. Future representations should continually incorporate new evidence, revise outdated conclusions, distinguish enduring characteristics from transient behaviors, preserve historical versions, and explicitly reason about uncertainty arising from incomplete or conflicting observations [30]. More fundamentally, future systems should not merely answer *"Who was this person?"* but continuously update their understanding of *"Who is this person now?"*

Person distillation also raises a deeper scientific question: What is the appropriate computational representation of a human? Existing approaches represent people using document-based representations, memory and knowledge systems, parametric models and adapters, preference and reward models, or hybrid representations [10, 37, 81, 135]. These are useful computational forms, but they are unlikely to capture the full richness of human expertise and judgment. Future work should investigate richer representations that preserve not only observable behaviors but also the principles, reasoning processes, causal relationships, and decision mechanisms that generate those behaviors. Such representations should remain interpretable, evidence-grounded, revisable, and capable of explaining *why* a person reached a particular conclusion rather than merely reproducing the conclusion itself. Addressing this challenge will require new connections among machine learning, knowledge representation, cognitive modeling, causal reasoning, and human cognition [96, 104]. Ultimately, person distillation may require new computational abstractions that bridge symbolic knowledge, statistical learning, episodic memory, and human reasoning into unified representations of people [66].

Another major frontier is extending person distillation from individuals to collective intelligence. Many forms of expertise reside not within individuals but within teams, organizations, and scientific communities [124]. A faithful collective representation should preserve complementary expertise, organizational memory, coordination procedures, authority structures, and legitimate disagreements rather than collapsing them into a synthetic "average" expert [11, 117]. Future AI systems may routinely consult multiple distilled people simultaneously, each contributing different expertise, assumptions, priorities, and historical perspectives. This raises new questions about routing problems to appropriate experts, reasoning over conflicting recommendations, preserving minority viewpoints, distinguishing disagreement caused by expertise from disagreement caused by outdated evidence, and determining when consensus should or should not be reached. In many settings, *preserving informed disagreement may be more faithful than producing artificial consensus*. Looking further ahead, person distillation may evolve into representations of laboratories, companies, governments, and scientific communities, enabling AI systems to preserve and reason over collective human knowledge rather than isolated individuals.

As person representations become increasingly capable, governance should become an integral component of their design rather than an external constraint. A distilled representation may continue operating long after its source traces become outdated, the represented person changes roles, withdraws consent, or is no longer available to supervise its use. Future systems therefore require mechanisms that place **people, rather than models, in control of**

their digital counterparts [24]. Consent should specify what information may be distilled, which aspects of a person may be represented, how the representation may be used, how long it may persist, and what level of authority it may exercise. Governance mechanisms should further support correction, provenance tracking, accountability, auditing, ownership, and transparent attribution of responsibility [9, 107]. Privacy presents an equally important challenge. Removing the original traces is insufficient if derived information remains embedded in memories, profiles, learned parameters, reward models, or collective representations. Future systems should therefore support **verifiable person unlearning**, allowing a person's influence to be identified, removed, and audited throughout the entire representation lifecycle [14, 33]. Designing person representations that remain simultaneously useful, controllable, accountable, and revocable will require close integration of machine learning, systems, security, and AI governance.

The ultimate goal is not autonomous person representations but trustworthy collaboration between humans and their computational counterparts. High person fidelity should not automatically imply unrestricted authority. A representation that faithfully captures someone's code-review decisions may provide little reliable guidance for hiring, strategic planning, or domains beyond the available evidence. Future systems should therefore communicate what they know, what they do not know, which facets are represented, which periods of a person's life are covered, and where important uncertainties remain. These limitations should directly determine what the system is permitted to do, ranging from answering questions to providing recommendations to making autonomous decisions. Future AI ecosystems may also require principled mechanisms for deciding *when to consult a distilled representation, when to seek additional expertise, when to defer to the real person if available, and when to abstain entirely* [31]. Trust calibration should therefore become an interactive property of long-term human–AI collaboration rather than a static confidence score attached to individual predictions [79].

Perhaps the most ambitious long-term challenge is moving beyond behavioral imitation toward causal and mechanistic fidelity. Existing systems primarily learn statistical regularities from historical traces. However, faithfully representing a person ultimately requires understanding *why* that person reaches particular conclusions, how competing objectives are balanced, what evidence changes their mind, and under what conditions they would make different decisions. Future evaluation should therefore move beyond reproducing historical behavior toward testing whether a representation responds appropriately to carefully controlled counterfactual situations [47, 48]. By systematically varying available evidence, organizational constraints, risks, stakeholders, or temporal context, future benchmarks may determine whether a representation captures the mechanisms underlying human judgment rather than merely its observable outcomes [32, 96]. Achieving causal fidelity would substantially expand the robustness, interpretability, and generalization capabilities of person-distillation systems, enabling them to reason responsibly beyond previously observed situations. More broadly, this direction suggests a transition from learning *what people did* to understanding *how people think*.

These directions suggest that **person distillation is evolving from a problem of constructing faithful representations into a broader scientific discipline concerned with modeling, maintaining, governing, and collaborating with computational representations of real people**. Achieving this vision will require advances in machine learning, knowledge representation, databases, systems, security, human–computer interaction, cognitive science, and AI governance. *The long-term goal is to enable AI systems that faithfully preserve, responsibly extend, and effectively collaborate with human expertise across individuals, organizations, and generations.* As personalized AI becomes increasingly collaborative, long-lived, and autonomous, we believe **person distillation has the potential to become a foundational abstraction connecting human knowledge with future intelligent systems**, much as knowledge representation, databases, and machine learning became foundational abstractions for previous generations of AI.

References

[1] 2026. Brother-Skill. https://github.com/realteamprinz/brother-skill. GitHub repository.

[2] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA, 1. doi:10.1145/1015330.1015430

[3] alchaincyf. 2026. Zhangxuefeng-Skill. https://github.com/alchaincyf/zhangxuefeng-skill. GitHub repository.

[4] Susan M Andersen and Serena Chen. 2002. The relational self: an interpersonal social-cognitive theory. *Psychological review* 109, 4 (2002), 619.

[5] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robot. Auton. Syst.* 57, 5 (May 2009), 469–483. doi:10.1016/j.robot.2008.10.024

[6] Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596.

[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[8] Sebastian Baltes and Stephan Diehl. 2018. Towards a theory of software development expertise. In *Proceedings of the 2018 26th acm joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 187–200.

[9] Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, et al. 2013. Prov-dm: The prov data model. *W3C Recommendation* 14 (2013), 15–16.

[10] Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. 2025. LoRe: Personalizing LLMs via Low-Rank Reward Modeling. arXiv:2504.14439 [cs.LG] https://arxiv.org/abs/2504.14439

[11] David P. Brandon and Andrea B. Hollingshead. 2004. Transactive Memory Systems in Organizations: Matching Tasks, Expertise, and People. *Organization Science* 15, 6 (2004), 633–644. doi:10.1287/orsc.1040.0069

[12] Susan Branje, Elisabeth L De Moor, Jenna Spitzer, and Andrik I Becht. 2021. Dynamics of identity development in adolescence: A decade in review. *Journal of Research on Adolescence* 31, 4 (2021), 908–927.

[13] Aylin Caliskan-Islam, Richard Harang, Andrew Liu, Arvind Narayanan, Clare Voss, Fabian Yamaguchi, and Rachel Greenstadt. 2015. De-anonymizing programmers via code stylometry. In *24th USENIX security symposium (USENIX Security 15)*. 255–270.

[14] Yinzhi Cao and Junfeng Yang. 2015. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*. 463–480. doi:10.1109/SP.2015.35

[15] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.

[16] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. 2024. PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences. arXiv:2406.08469 [cs.LG] https://arxiv.org/abs/2406.08469

[17] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: reinforcement learning via sequence modeling. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1156, 14 pages.

[18] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[19] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413* (2025).

[20] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[21] Brian S Connelly and Deniz S Ones. 2010. An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. *Psychological bulletin* 136, 6 (2010), 1092.

[22] Ziyao Cui, Minxing Zhang, and Jian Pei. 2026. On Membership Inference Attacks in Knowledge Distillation. *SIGKDD Explor. Newsl.* 28, 1 (June 2026), 32–40. doi:10.1145/3820356.3820359

[23] dadwadw233. 2026. VibePortrait. https://github.com/dadwadw233/VibePortrait. GitHub repository.

[24] John Danaher and Sven Nyholm. 2025. The ethics of personalised digital duplicates: a minimally viable permissibility principle. *AI and Ethics* 5, 2 (2025), 1703–1718. doi:10.1007/s43681-024-00513-7

[25] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems* 36 (2023), 10088–10115.

[26] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[27] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Commun. ACM* 67, 1 (2023), 110–120.

[28] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological review* 100, 3 (1993), 363.

[29] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[30] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.

[31] Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems* 30 (2017).

[32] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research* 26, 83 (2025), 1–64.

[33] Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. 2019. Making AI forget you: data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 316, 14 pages.

[34] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision* 129, 6 (2021), 1789–1819.

[35] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Your behavior signals your reliability: Modeling crowd behavioral traces to ensure quality relevance annotations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 41–49.

[36] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[37] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.

[38] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. 263–272. doi:10.1109/ICDM.2008.22

[39] Rob J Hyndman and George Athanasopoulos. 2018. *Forecasting: principles and practice*. OTexts.

[40] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[41] Yanna Jiang, Delong Li, Haiyu Deng, Baihe Ma, Xu Wang, Qin Wang, and Guangsheng Yu. 2026. SoK: Agentic Skills–Beyond Tool Use in LLM Agents. *arXiv preprint arXiv:2602.20867* (2026).

[42] jiangziyan-693. 2026. MamaSkill. https://github.com/jiangziyan-693/MamaSkill. GitHub repository.

[43] Mitchell Joblin, Sven Apel, and Wolfgang Mauerer. 2017. Evolutionary trends of developer coordination: A network approach. *Empirical Software Engineering* 22, 4 (2017), 2050–2094.

[44] Daniel Kahneman and Gary Klein. 2009. Conditions for intuitive expertise: a failure to disagree. *American psychologist* 64, 6 (2009), 515.

[45] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[46] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 4 (2012), 1–21.

[47] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the Difference that Makes a Difference with Counterfactually Augmented Data. *International Conference on Learning Representations (ICLR)* (2020).

[48] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4466–4474. Survey Track. doi:10.24963/ijcai.2021/609

[49] Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 465–470.

[50] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.

[51] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45, 1 (2011), 83–94.

[52] Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the association for information science and technology* 65, 1 (2014), 178–187.

[53] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 447–456.

[54] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.

[55] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[56] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597* (2023).

[57] Mengdi Li, Guanqiao Chen, Xufeng Zhao, Haochen Wen, Shu Yang, and Di Wang. 2025. PersRM-R1: Enhance Personalized Reward Modeling with Reinforcement Learning. arXiv:2508.14076 [cs.LG] https://arxiv.org/abs/2508.14076

[58] Paul Luo Li, Amy J Ko, and Jiamin Zhu. 2015. What makes a great software engineer?. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 700–710.

[59] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).

[60] Shangsong Liang. 2019. Collaborative, dynamic and diversified user profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4269–4276.

[61] Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. Llms+ persona-plug= personalized llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9373–9385.

[62] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. doi:10.18653/v1/2023.emnlp-main.153

[63] Dylan P. Losey and Marcia K. O'Malley. 2018. Including Uncertainty when Learning from Human Corrections. In *Proceedings of The 2nd Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 87)*, Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto (Eds.). PMLR, 123–132. https://proceedings.mlr.press/v87/losey18a.html

[64] Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems* 37 (2024), 74325–74362.

[65] Heidy M. Marin-Castro and Edgar Tello-Leal. 2021. Event Log Preprocessing for Process Mining: A Review. *Applied Sciences* 11, 22 (2021). doi:10.3390/app112210556

[66] Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. 2024. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence* 328 (2024), 104062. doi:10.1016/j.artint.2023.104062

[67] Robert R McCrae and Paul T Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist* 52, 5 (1997), 509.

[68] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[69] Shaunak A. Mehta and Dylan P. Losey. 2024. Unified Learning from Demonstrations, Corrections, and Preferences during Physical Human–Robot Interaction. *J. Hum.-Robot Interact.* 13, 3, Article 39 (Aug. 2024), 25 pages. doi:10.1145/3623384

[70] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems* 35 (2022), 17359–17372.

[71] Taciano L Milfont, Petar Milojev, and Chris G Sibley. 2016. Values stability and change in adulthood: A 3-year longitudinal study of rank-order stability and mean-level differences. *Personality and Social Psychology Bulletin* 42, 5 (2016), 572–588.

[72] Michael E. Miller and Emily Spatz. 2022. A unified view of a human digital twin. *Human-Intelligent Systems Integration* 4, 1 (2022), 23–33. doi:10.1007/s42454-022-00041-x

[73] Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review* 102, 2 (1995), 246.

[74] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309* (2021).

[75] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)* 50, 6 (2017), 1–36.

[76] Andrew Y. Ng and Stuart J. Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 663–670.

[77] Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. 2026. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158* (2026).

[78] Ikujiro Nonaka. 1994. A dynamic theory of organizational knowledge creation. *Organization science* 5, 1 (1994), 14–37.

[79] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (02 2020), 1–20. doi:10.1371/journal.pone.0229132

[80] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[81] Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. MemGPT: towards LLMs as operating systems. (2023).

[82] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[83] James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77, 6 (1999), 1296.

[84] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*. 487–503.

[85] Michael Polanyi. 2009. The tacit dimension. In *Knowledge in organisations*. Routledge, 135–146.

[86] Adithya Pratapa and Teruko Mitamura. 2025. Scaling Multi-Document Event Summarization: Evaluating Compression vs. Full-Text Approaches. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 514–528.

[87] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2338, 14 pages.

[88] Rafael A Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 913–919.

[89] Brent W Roberts, Kate E Walton, and Wolfgang Viechtbauer. 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin* 132, 1 (2006), 1.

[90] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 15)*, Geoffrey Gordon, David Dunson, and Miroslav Dudík (Eds.). PMLR, Fort Lauderdale, FL, USA, 627–635. https://proceedings.mlr.press/v15/ross11a.html

[91] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[92] A. Rozinat and W.M.P. {Aalst, van der}. 2006. *Decision mining in business processes*. Technische Universiteit Eindhoven.

[93] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927* 1 (2024).

[94] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7370–7392.

[95] Francesco Sanna Passino, Lucas Maystre, Dmitrii Moor, Ashton Anderson, and Mounia Lalmas. 2021. Where To Next? A Dynamic Model of User Preferences. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3210–3220. doi:10.1145/3442381.3450028

[96] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward Causal Representation Learning. *Proc. IEEE* 109, 5 (2021), 612–634. doi:10.1109/JPROC.2021.3058954

[97] Burr Settles. 2009. Active learning literature survey. (2009).

[98] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13153–13187.

[99] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).

[100] Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. 2025. Language Model Personalization via Reward Factorization. arXiv:2503.06358 [cs.LG] https://arxiv.org/abs/2503.06358

[101] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. 2020. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences* 117, 30 (2020), 17680–17687.

[102] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60, 3 (2009), 538–556.

[103] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 253, 14 pages.

[104] Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427* (2023).

[105] Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112, 1 (1999), 181–211. doi:10.1016/S0004-3702(99)00052-1

[106] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* 10, 05 (2002), 557–570.

[107] Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). doi:10.6028/NIST.AI.100-1

[108] Niek Tax, Natalia Sidorova, Reinder Haakma, and Wil M. P. van der Aalst. 2018. Event Abstraction for Process Mining Using Supervised Learning Techniques. In *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*, Yaxin Bi, Supriya Kapoor, and Rahul Bhatia (Eds.). Springer

International Publishing, Cham, 251–269.

[109] titanwings. 2026. Ex-Skill. https://github.com/titanwings/ex-skill. GitHub repository.

[110] Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11836–11850. doi:10.18653/v1/2024.acl-long.638

[111] Wil Van Der Aalst. 2016. Data science in action. In *Process mining: Data science in action*. Springer, 3–23.

[112] Wil van der Aalst, Arya Adriansyah, Ana Karla Alves de Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter van den Brand, Ronald Brandtjen, Joos Buijs, Andrea Burattin, Josep Carmona, Malu Castellanos, Jan Claes, Jonathan Cook, Nicola Costantini, Francisco Curbera, Ernesto Damiani, Massimiliano de Leoni, Pavlos Delias, Boudewijn F. van Dongen, Marlon Dumas, Schahram Dustdar, Dirk Fahland, Diogo R. Ferreira, Walid Gaaloul, Frank van Geffen, Sukriti Goel, Christian Günther, Antonella Guzzo, Paul Harmon, Arthur ter Hofstede, John Hoogland, Jon Espen Ingvaldsen, Koki Kato, Rudolf Kuhn, Akhil Kumar, Marcello La Rosa, Fabrizio Maggi, Donato Malerba, Ronny S. Mans, Alberto Manuel, Martin McCreesh, Paola Mello, Jan Mendling, Marco Montali, Hamid R. Motahari-Nezhad, Michael zur Muehlen, Jorge Munoz-Gama, Luigi Pontieri, Joel Ribeiro, Anne Rozinat, Hugo Seguel Pérez, Ricardo Seguel Pérez, Marcos Sepúlveda, Jim Sinur, Pnina Soffer, Minseok Song, Alessandro Sperduti, Giovanni Stilo, Casper Stoel, Keith Swenson, Maurizio Talamo, Wei Tan, Chris Turner, Jan Vanthienen, George Varvaressos, Eric Verbeek, Marc Verdonk, Roberto Vigo, Jianmin Wang, Barbara Weber, Matthias Weidlich, Ton Weijters, Lijie Wen, Michael Westergaard, and Moe Wynn. 2012. Process Mining Manifesto. In *Business Process Management Workshops*, Florian Daniel, Kamel Barkaoui, and Schahram Dustdar (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 169–194.

[113] Sebastiaan J. van Zelst, Felix Mannhardt, Massimiliano de Leoni, and Agnes Koschmider. 2021. Event abstraction in process mining: literature review and taxonomy. *Granular Computing* 6, 3 (2021), 719–736. doi:10.1007/s41066-020-00226-2

[114] Simine Vazire. 2010. Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of personality and social psychology* 98, 2 (2010), 281.

[115] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796* (2024).

[116] vogtsw. 2026. Boss-Skills. https://github.com/vogtsw/boss-skills. GitHub repository.

[117] James P. Walsh and Gerardo Rivera Ungson. 1991. Organizational Memory. *The Academy of Management Review* 16, 1 (1991), 57–91. doi:10.5465/amr.1991.4278992

[118] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

[119] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. arXiv:2406.12845 [cs.LG] https://arxiv.org/abs/2406.12845

[120] Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. 14743–14777.

[121] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9440–9450.

[122] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *Comput. Surveys* 57, 3 (2024), 1–37.

[123] Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. COSER: coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning* (Vancouver, Canada) *(ICML'25)*. JMLR.org, Article 2576, 37 pages.

[124] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (2010), 686–688. doi:10.1126/science.1193147

[125] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. arXiv:2306.01693 [cs.CL] https://arxiv.org/abs/2306.01693

[126] xixu-me. 2026. Awesome Persona Distill Skills. https://github.com/xixu-me/awesome-persona-distill-skills. GitHub repository, commit 37a2e85, accessed July 15, 2026.

[127] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems* 35 (2022), 20744–20757.

[128] ybq22. 2026. Mentor.skill. https://github.com/ybq22/supervisor. GitHub repository.

[129] Ziqin Yuan, Ruiqi Wang, Dezhong Zhao, Baijian Yang, and Byung-Cheol Min. 2026. PrefMoE: Robust Preference Modeling with Mixture-of-Experts Reward Learning. arXiv:2605.00384 [cs.RO] https://arxiv.org/abs/2605.00384

[130] Minxing Zhang, Yi Yang, Zhuofan Jia, Xuan Yang, Jian Pei, Yuchen Zang, Xingwang Deng, and Xianglong Chen. 2026. MPCEval: A Benchmark for Multi-Party Conversation Generation. *arXiv preprint arXiv:2603.04969* (2026).

[131] Minxing Zhang, Yi Yang, Roy Xie, Bhuwan Dhingra, Shuyan Zhou, and Jian Pei. 2026. Generalizability of Large Language Model-Based Agents: A Comprehensive Survey. *ACM Comput. Surv.* 58, 10, Article 263 (April 2026), 44 pages. doi:10.1145/3794858

[132] 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)*. 2204–2213.

[133] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*. Pmlr, 12697–12706.

[134] Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. CHARACTERBENCH: benchmarking character customization of large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 2908, 10 pages. doi:10.1609/aaai.v39i24.34806

[135] Tianyi Zhou, Dongrui Liu, Leitao Yuan, Jing Shao, and Xia Hu. 2026. COLLEAGUE.SKILL: Automated AI Skill Generation via Expert Knowledge Distillation. arXiv:2605.31264 [cs.AI] https://arxiv.org/abs/2605.31264

[136] Thomas Zollo, Andrew Siah, Naimeng Ye, Li Li, and Hongseok Namkoong. 2025. Personalllm: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, Vol. 2025. 66949–66971.