

Creating an Atomic User Model for Personality-Aware Large Language Model Interaction

B. Sankar^a, Deepthika S^b, Pawni Yadav^b, Amogh A S^b

^a*Department of Mechanical Engineering, Indian Institute of Science (IISc), Bengaluru, 560012, Karnataka, India*

^b*Department of Design and Manufacturing, Indian Institute of Science (IISc), Bengaluru, 560012, Karnataka, India*

^c*Corresponding Author: B. Sankar*

Corresponding Author Email: sankarb@iisc.ac.in,

Abstract

Assistants built on large language models are increasingly expected to write as their user would write, and the dominant approach to that expectation operates on a single channel: past preferences are summarised out of conversation history and reinserted into the context window. We argue that this inverts the natural order of inference. Preferences are the task-dependent surface of an underlying personality structure that is comparatively stable, so a system that stores only preferences must relearn the person whenever the task changes. This paper makes four contributions. First, we characterise a phenomenon we call *personality seepage*, in which the linguistic surface of a prompt carries a personality fingerprint that the assistant mirrors without any access to the personality that produced it. Second, we propose the Atomic User Model (AUM), a structured, human-readable representation that organises a person as a stable identity nucleus surrounded by four interpretable shells covering psychological, cognitive and experiential, behavioural, and social content, together with cross-shell entries recording internal conflict and authenticity. Third, we treat AUM as a retrieval index over a person rather than a prompt prefix, and specify a personality-aware pipeline in which a task classifier, a component-selection function and a budgeted retriever return a small payload of fields at generation time. Fourth, we evaluate the pipeline in a simulation study with sixteen language-model-simulated participants, six style-sensitive tasks and three seeds, and in a synthetic scaling study of the retriever itself. Retrieving eight fields matched the style fidelity of injecting the entire user model while using 23% of the context (211 tokens against 915), improved on flat preference notes by 0.24 points on a five-point scale ($p < 0.001$, $d_z = 0.50$), and raised forced-choice identification of the participant’s own voice from 14.9% to 42.7% against 25% chance. Four pre-registered controls returned null, including the swarm retriever against forward greedy, which locates the effect in the structured representation rather than in the search over it. A supplementary analysis shows the benefit is largest for participants whose voice the un-personalised assistant reproduces worst ($\rho = -0.61$, $p = 0.013$), which is a design result rather than a statistical one: personalisation is worth most to the people the default serves least. We discuss what a budgeted payload buys for privacy-by-architecture, and what a simulation can and cannot establish about whether real people recognise themselves.

Keywords: user modelling, personalisation, personality, large language models, retrieval augmented generation, context budget, scrutability, privacy by architecture, simulation study, artificial fish swarm algorithm

1. Introduction

Adaptive systems have modelled their users for as long as they have adapted to them. The generic user modelling tradition in human-computer interaction established that a user model is a reusable component with an explicit schema, maintained separately from the application that consumes it (Kobsa, 2001; Brusilovsky and Millán, 2007), and personalised search built profiles from observed behaviour: interests and activities harvested from a desktop index (Teevan et al., 2005), click and query histories evaluated at scale (Dou et al., 2007), and short-term session signals combined with long-term interest (Bennett et al., 2012). Assistants built on large language models (LLMs) have inherited this design in a compressed form. Deployed memory features distil past conversations into short preference notes and reinsert them into later contexts, and the research literature has followed with memory architectures (Packer et al., 2023; Zhou et al., 2024; Liao et al., 2026), per-user parameter-efficient adaptation (Tan et al., 2024), and surveys that map the resulting space (Tan and Jiang, 2023; Zhang et al., 2025).

What these approaches share is not an algorithm but a unit of storage. They store what a user has done, or has said they want. They do not store the person from whom those wants emerge. The difference is easiest to see when a stored preference goes stale.

A user plans a six-day conference trip to Paris and discusses cost of living and itineraries with an assistant. Two weeks later the user privately decides not to travel, and does not say so. A month after that, while drafting a project timeline, the assistant blocks out the original travel week as an absence.

The system is not wrong about its source: it remembers the preference faithfully, and the retrieved item was exactly the right item. It is wrong about the person. It holds no representation of the user as a decision maker whose plans get revised, so it has no basis on which to flag a stale assumption. Retrieval quality is not the failure here, and better retrieval over the same unit would not have prevented it.

This paper argues that the missing layer is a different unit of storage. The argument rests on a claim borrowed from personality psychology rather than information retrieval. Personality traits show substantial rank-order stability across the lifespan, with test-retest correlations rising from roughly 0.3 in childhood to 0.7 in later adulthood (Roberts and DelVecchio, 2000), and mean-level change, while real, is gradual and patterned (Roberts et al., 2006; Specht et al., 2011). Preferences, by contrast, are situational: they are what personality produces when it meets a particular decision context. If the stable layer is the one that generalises across tasks, then a system that stores only the unstable layer is storing the wrong thing.

We make four contributions.

1. **Personality seepage.** We characterise a phenomenon in which a user’s linguistic and behavioural patterns enter the assistant through the surface of their prompt, and the assistant mirrors that surface while having no access to the personality that produced it (Section 3).

*Corresponding author: B. Sankar

**Corresponding author email: sankarb@iisc.ac.in

Email addresses: sankarb@iisc.ac.in (B. Sankar), deepthikas123@gmail.com (Deepthika S), pawniyadav3435@gmail.com (Pawni Yadav), amogh.setty07@gmail.com (Amogh A S)

URL: 0000-0001-5844-6273 (B. Sankar), 0009-0009-4386-6510 (Deepthika S)

Seepage is why generic assistants can appear briefly personalised and then fail on exactly the parts of a task the user did not spell out.

2. **The Atomic User Model.** We propose a structured, human-readable representation that organises a person as a stable identity Nucleus surrounded by four shells covering psychological, cognitive and experiential, behavioural, and social content, plus cross-shell entries recording internal conflict, growth trajectory and an authenticity index (Section 4). The metaphor is not decorative: it encodes two design commitments, differential stability and differential observability, that a flat schema cannot express.
3. **Personality-aware retrieval.** We treat AUM as an index over a person rather than a prompt prefix, and specify a pipeline in which a task classifier, a component-selection function $\sigma(t)$ and a budgeted retriever return a payload of k fields at generation time (Section 5). The retrieval objective carries an explicit redundancy penalty, which is what makes payload selection a combinatorial problem rather than a sort.
4. **Two studies.** We evaluate the pipeline with sixteen LLM-simulated participants across six style-sensitive tasks and three seeds (Sections 6 and 7), and we study the retrieval problem on its own in a synthetic sweep over instrument size, budget and redundancy weight (Section 8).

The headline empirical result is a context result. Eight retrieved fields matched the style fidelity of injecting the entire 32-field user model while using 23% of the injected context, and improved on flat preference notes by 0.24 points on a five-point scale. Forced-choice identification of the participant’s own voice rose from 14.9% under preference notes, which is not distinguishable from chance, to 42.7% against a 25% baseline.

Four controls returned null, and we report them in full because together they bound the claim. Selecting the right k fields was not distinguishable from selecting k at random from the same admissible pool; component selection contributed nothing beyond retrieval; the swarm retriever neither beat forward greedy on generated style nor reached a higher value of its own objective; and the redundancy term did not earn its place against a relevance-only top- k . The scaling study of Section 8 explains why, and the explanation is specific: at the operating point the simulation used, forward greedy is already within 0.1% of the exhaustive optimum, so there was nothing for a population method to recover. What the positive results measure is the structured representation, not the search over it.

A supplementary analysis produced the finding we think matters most for a human-computer interaction audience. The benefit of personalisation was largest exactly for those participants whose voice the un-personalised assistant reproduced worst (Spearman $\rho = -0.61$, $p = 0.013$). Personalisation is worth most to the people the default serves least, which is a statement about who a system is for rather than about how well it scores.

Figure 1 states the inversion in one picture.

The remainder of the paper is organised as follows. Section 2 situates the work. Section 3 defines personality, preferences and persona and introduces seepage. Section 4 specifies AUM. Section 5 formalises personality-aware retrieval. Sections 6 and 7 present the simulation study and its results. Section 8 reports the scaling study. Section 9 discusses implications and limitations, and Section 11 concludes. Appendices give the full specification, every prompt used, the persona set and the complete statistical tables.

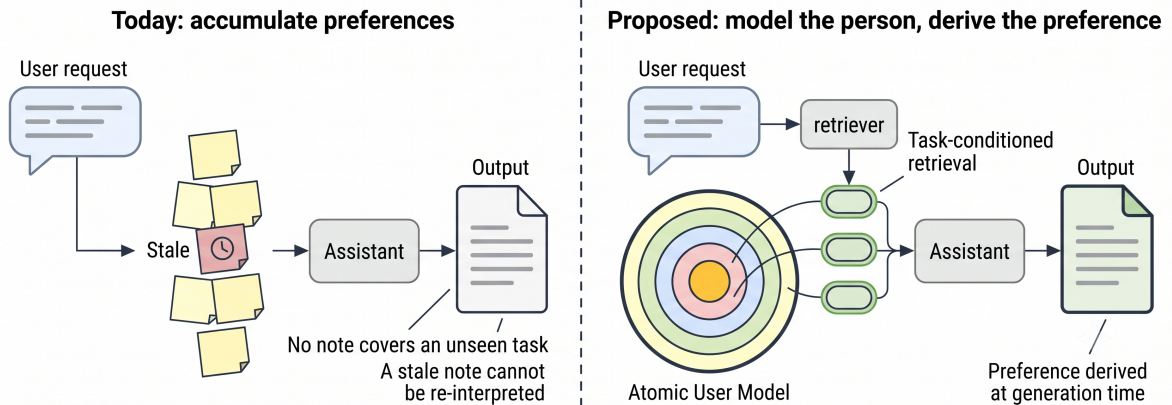


Figure 1: The inversion this paper proposes. Left: current personalisation accumulates preferences observed at the surface and reinserts them, so an unseen task has no covering note and a stale note cannot be re-interpreted. Right: AUM stores the comparatively stable personality structure and derives the situational preference at generation time, retrieving only the fields the task needs.

2. Related Work

2.1. User models as components in adaptive systems

The idea that a user model should be a separable, inspectable component predates the systems this paper is about. Kobsa (2001) surveyed generic user modelling systems and argued for a shared server holding assumptions about the user, queried by applications rather than duplicated inside them. Brusilovsky and Millán (2007) developed the same position for adaptive hypermedia and educational systems, where an overlay model over a domain structure gives the adaptation something explicit to reason about. Information retrieval developed a parallel but distinct tradition, in which the user model is a statistical summary of behaviour rather than a schema: term vectors and topical distributions built from a desktop index (Teevan et al., 2005), click and query histories whose value for ranking turns out to vary sharply across queries (Dou et al., 2007), and a decomposition of behavioural evidence into short-term session signals and long-term interest (Bennett et al., 2012). Diversity-aware ranking developed alongside, first as maximal marginal relevance (Carbonell and Goldstein, 1998), then as an explicit evaluation concern (Clarke et al., 2008) and an optimisation target (Agrawal et al., 2009).

AUM belongs to the first tradition and borrows from the second. It is a schema, in Kobsa’s sense, whose slots are declared in advance rather than induced from data. What it takes from information retrieval is the insight that a profile is only useful to the extent that a system can select from it, and that selecting well means trading relevance against redundancy.

2.2. Memory and personalisation in LLM assistants

Recent work recasts personalisation of LLM assistants as memory management. MemGPT treats the model as an operating system with paged memory (Packer et al., 2023); generative agents combine memory, reflection and planning to produce believable behaviour (Park et al., 2023); medical and lifelong assistants coordinate short-term and long-term stores (Zhang et al., 2024a; Wang et al., 2024b). Closer to retrieval, cognitive personalised search couples an LLM with

an efficient memory mechanism over query logs (Zhou et al., 2024), profile memory organises per-user summaries for reuse (Liao et al., 2026), and agent-driven retrieval personalises search results directly (Chhetri et al., 2026). A separate branch adapts model parameters per user rather than the context, through per-user parameter-efficient fine-tuning (Tan et al., 2024); a recent survey maps the whole space (Zhang et al., 2025).

Evaluation has developed in step. LaMP (Salemi et al., 2024) and its long-form extension (Kumar et al., 2024) define personalisation as task-level retrieval over user history. PersonaBench shows that retrieval-augmented systems answer questions about private user data poorly even when given direct document access (Tan et al., 2025), and rating-prediction studies find that LLMs capture user preferences but in opaque ways (Kang et al., 2023; Huang et al., 2025).

These systems answer how to store and fetch user-relevant content. AUM answers a prior question: what the stored items should be, and how they should be organised so that fetching a few of them is enough. The two are complementary, and the pipeline of Section 5 is deliberately built from standard components so that the representation, rather than the machinery, is what is under test.

2.3. Context budgets and compression

A structured user model of sixty fields cannot be injected on every turn. The problem of fitting useful conditioning into a bounded context has been attacked from the compression side: LLM-Lingua and its long-context successor learn to drop tokens from a prompt while preserving downstream performance (Jiang et al., 2023, 2024), and selective-context methods prune low-information spans (Li et al., 2023). Our approach is complementary and operates a level up. Rather than compressing a fixed payload, we select which fields enter the payload at all, conditioned on the task. The two could compose: compression could shorten the eight fields that selection returns.

2.4. Personality: structure, stability, and its trace in language

The Five-Factor Model provides the trait vocabulary most widely used in computational work (McCrae and John, 1992), with modern instruments including the BFI-2 (Soto and John, 2017), public-domain item pools (Goldberg et al., 2006) and very brief measures for settings where length is prohibitive (Gosling et al., 2003). Cross-cultural generalisation of the factor structure has been examined but remains a live concern (McCrae, 2002). We include Myers-Briggs type as a field inside Shell 1 because users often know their type, while noting that its psychometric properties are weak: the type dichotomies do not correspond to natural categories, and reliability across retests is poor (Pittenger, 2005; Stein and Swan, 2019; McCrae and Costa, 1989). This is precisely the argument for a schema in which an inventory is one field among many rather than the representation itself.

That personality leaves a measurable trace in language is well established. Pennebaker and King (1999) showed that function-word use is a stable individual difference; Yarkoni (2010) related Big Five scores to word use across 100,000 words of blog text; Schwartz et al. (2013) moved from closed-vocabulary counting to open-vocabulary differential language analysis over social media, and the LIWC line of work supplies the standard closed-vocabulary instrument (Tausczik and Pennebaker, 2010; Boyd and Pennebaker, 2017). Prediction of traits from digital footprints reaches useful accuracy (Kosinski et al., 2013; Park et al., 2015), sometimes exceeding human judges given enough behavioural evidence (Youyou et al., 2015), though meta-analysis places the typical correlation more modestly (Azucar et al., 2018), and earlier attempts to model personality from social media profile data illustrate how far the field has moved (Goyal and Tawde, 2022). Within LLM research, trait recognition from text is feasible (Ji et al., 2023; Rao et al., 2023) and implicit persona can be

modelled as a latent variable in dialogue (Cho et al., 2022), but the resulting representations are trait scores or latent vectors: neither modular nor inspectable.

The reliance on self-report in Section 6 inherits a known limitation. Self and other observers know different things about a person, with the self better on internal states and observers better on externally visible traits (Vazire, 2010; Connelly and Ones, 2010), and self-report has long been criticised as a substitute for observed behaviour (Baumeister et al., 2007). We return to this in Section 9.5.

2.5. Structured user representations and human digital twins

The closest contemporary architecture is the General User Model, which accumulates confidence-weighted natural-language propositions about a user from screen observations (Shaikh et al., 2025). Generative agents grounded in self-report interviews simulate individual humans better than demographic conditioning alone (Park et al., 2024), which is both the closest precedent for our representation and the precedent for our evaluation method. Both populate a representation bottom up, from observation or interview. AUM is top down: it specifies the slots into which such observations should be organised, which is what makes a component-selection function over shells definable at all.

The human digital twin literature pursues a superficially similar goal from an engineering direction, and a recent survey maps its scope (Lin et al., 2024). The personal informatics tradition in HCI supplies the complementary user-centred account of what it is like to live with a model of oneself: staged models of collection and reflection (Li et al., 2010), the lived-informatics revision that treats tracking as episodic rather than linear (Epstein et al., 2015), and evidence on how people without prior self-tracking experience engage with their own data (Rapp and Cena, 2016). AUM differs from a digital twin in purpose. It is built for style-sensitive task completion, not identity substitution, and Section 9 argues that the distinction has consequences the twin framing obscures.

2.6. Scrutability, control, and the privacy of a user model

If a system holds a model of a person, the person should be able to read it. Kay and Kummerfeld (2012) set out the drivers and principles for personalised systems that users can scrutinise and control, building on earlier work on scrutable adaptive hypertext (Czarkowski and Kay, 2002) and on interfaces for inspecting semantic user models (Bakalov et al., 2010). Cramer et al. (2008) found that transparency affected acceptance of a content-based recommender, and Harper et al. (2015) showed that giving users direct control over their recommendations changed both behaviour and satisfaction. Jeromela (2022) extends the question to intelligent personal assistants specifically. Mental-model work supplies the reason this matters: users arrive with expectations of conversational agents that the systems do not meet (Luger and Sellen, 2016; Mahmood et al., 2025), and their models of what an agent knows shape how they use it (Gero et al., 2020; Ngo et al., 2020).

The privacy dimension is sharper for AUM than for a preference list, because the content is more intimate. Contextual integrity supplies the governing frame: what matters is not secrecy but whether an information flow matches the norms of the context it came from (Nissenbaum, 2011). Empirically, users of LLM-based conversational agents disclose a great deal and reason about the risks only partially (Zhang et al., 2024b), and language models can leak personal information present in training data (Huang et al., 2022). On-device and federated approaches offer one architectural answer (Chen et al., 2019). Our position, developed in Sections 5 and 10, is that a budgeted payload converts local residence from a policy promise into a measurable property: if eight fields suffice, the other twenty-four never leave the device.

2.7. Writing assistance and the user’s voice

The application setting for this work is writing assistance, where the question of whose voice appears in the output is not incidental. CoAuthor documented human-LLM collaborative writing at scale (Lee et al., 2022); Wordcraft explored story writing with a model in the loop (Yuan et al., 2022); a recent design space synthesises the field (Lee et al., 2024) and a companion study aligns research directions with what writers actually ask for (Reza et al., 2025). The finding that most directly motivates AUM is Jakesch et al. (2023): co-writing with an opinionated language model shifts the user’s own expressed views. If a default assistant voice can move what a person says, then a system that instead conditions on a model of that person is not merely a convenience feature.

2.8. Selecting a subset under a budget

Choosing k fields to maximise a payload score with a redundancy penalty is a constrained subset-selection problem. When such an objective is monotone submodular, the forward greedy algorithm carries the classical $(1 - 1/e)$ guarantee (Nemhauser et al., 1978), which is why greedy is the right baseline rather than a straw man. Our objective, defined in Section 5, uses a mean rather than a sum and subtracts a mean pairwise similarity, so it is not submodular in general and the guarantee does not transfer; the empirical question of how close greedy comes is therefore live, and Section 8 answers it directly.

Population-based metaheuristics are widely applied to subset selection, particularly feature selection (Xue et al., 2016; Nguyen et al., 2020). The artificial fish swarm algorithm (Li et al., 2002) models a population of agents executing prey, swarm and follow behaviours, and has accumulated a substantial literature of variants and applications (Neshat et al., 2014; Pourpanah et al., 2023). We use it, and we report that at the scale our simulation ran it did not earn its cost. We think reporting that is more useful than dropping the comparison, and Section 8 identifies the regime in which the answer changes.

2.9. Simulated participants and model-based judgement

Using language models in place of human participants is now a small field with a sharp internal debate. Argyle et al. (2023) showed that conditioning on demographic backstories reproduces aggregate patterns in survey data, and Park et al. (2024) showed that grounding agents in structured self-report improves individual-level simulation. The critical literature is equally developed: Bisbee et al. (2024) document instability and misestimation in synthetic survey responses, and Wang et al. (2025) show that replacing human participants can misportray and flatten identity groups. We adopt simulation as a feasibility method with explicit limits, stated in Section 6.1, and we do not present any result here as evidence about human users.

Model-based judgement carries its own hazards, all of which we control for by construction. Judges exhibit position bias (Wang et al., 2024a), self-preference for their own generations (Panickssery et al., 2024), and imperfect but usable agreement with human raters (Zheng et al., 2023; Chiang and Lee, 2023; Liu et al., 2023). Section 6.7 describes the specific controls: a judge from a different model family than the generator, per-trial option shuffling, and a judge that never sees the user model it is implicitly evaluating.

For the automatic style measure we use authorship representations rather than semantic embeddings. LUAR learns universal authorship representations by contrastive training over authors (Rivera-Soto et al., 2021), extending earlier invariant representations of social-media users (Andrews and Bishop, 2019). The distinction matters here: a semantic embedding scores any two

deadline-extension emails as similar because they are about the same thing, which is exactly the wrong invariance. Work on what authorship representations actually encode (Wang et al., 2023; Wegmann et al., 2022), on making style embeddings interpretable (Patel et al., 2023), and on robustness (Man and Nguyen, 2024) informs how we read the resulting numbers. Style-transfer evaluation supplies the broader methodological caution that automatic style metrics are easy to misreport (Mir et al., 2019; Ostheimer et al., 2023), and fine-grained linguistic control offers an alternative route to the same goal (Alhafni et al., 2024).

2.10. Cognitive architectures

Finally, the design philosophy behind a fixed, modular schema comes from the cognitive architecture tradition. ACT-R (Anderson et al., 2004) and Soar (Laird, 2012) are long-standing arguments that explicit structure buys something that implicit competence does not, and Sun (2024) makes the case that such architectures and LLMs are complementary rather than competing. AUM applies the philosophy not to cognition in general but to the narrower problem of representing the user inside an LLM-mediated system.

3. Personality, Preferences, Persona, and Seepage

3.1. Three terms used in a specific way

Work on personalisation uses *personality*, *preference* and *persona* loosely and often interchangeably. Because the argument of this paper is precisely about the relation between them, we fix the terms.

Personality is the structured set of stable, task-independent traits spanning the cognitive, affective, behavioural and social-contextual dimensions of a person. Stability is meant in the rank-order sense established empirically by Roberts and DelVecchio (2000): a person’s position relative to others changes slowly, even where absolute levels drift with age and life events (Roberts et al., 2006; Specht et al., 2011).

Preferences are the task-dependent outputs of personality applied to a particular decision context. “Prefers bullet points in status updates” is a preference. It is downstream of conscientiousness, of a professional register learned in a particular workplace, and of a belief about what respects a reader’s time; it is not itself any of those things.

Persona is the externally perceived projection of personality through one channel: a professional profile, a social media account, a conference biography. Goffman’s account of self-presentation (Goffman, 1959) is the reference point, and the key property is that a persona is a selective and audience-dependent rendering rather than a compressed copy.

AUM models personality. Preferences are derived from personality and context at generation time. Personas are situational projections and are represented, when they are represented at all, as fields inside Shell 4.

The ordering matters for a practical reason. If preferences are the stored unit, then every new task requires either a stored preference that happens to cover it or a guess. If personality is the stored unit, a preference for an unseen task is derivable, and a stale stored preference can be re-interpreted against a structure that did not change. The Paris example of Section 1 is exactly a case where the second operation was needed and unavailable.

Table 1: Personality seepage on a single task: an email requesting a deadline extension from a colleague. Three users phrase the same underlying request differently and the assistant mirrors each surface. The third column is the point: what the generic response misses is in each case a property of the user, not of the prompt.

User	Prompt surface (abridged)	What the generic response misses
U1: organised, meticulous	“Current deadline Friday 5pm; requested Monday 5pm. Reason: data validation overran scope. Offer: draft by Saturday.”	The structure is right and the dates are carried through, but the reply drops the acknowledgement of the reviewer’s effort that U1 includes in every message they send. Nothing in the prompt asks for it.
U2: empathetic, verbose	“Could you help me write a friendly note to a colleague? I feel a bit bad asking, but maybe we could push the Q3 deadline.”	The reply amplifies the apology into self-deprecation. U2’s phrasing invites it; U2’s stated values reject it. Mirroring the surface overshoots the person.
U3: concise, informal	“email to ask for more time on a report, colleague, Q3 stuff. make it not too formal.”	Brevity is preserved and so is the omission: no date and no reason, the two things U3 habitually leaves out of a first draft and reliably needs by the second.

3.2. Personality seepage

Table 1 illustrates a phenomenon we call *personality seepage*. When a person writes a prompt, their linguistic and behavioural patterns go into the prompt with them, and the model amplifies those patterns in its output. That the trace is there is not in doubt: function-word use is a stable individual difference (Pennebaker and King, 1999), and personality is recoverable from text at useful accuracy (Yarkoni, 2010; Schwartz et al., 2013).

The trouble is what the assistant does with it. It matches the surface without access to what produced the surface. Three consequences follow, and each appears in Table 1.

1. **Gaps are not filled.** Whatever the user habitually omits from a prompt is also absent from the output, because the model has no independent source for it. U3’s missing deadline is the example.
2. **Surface features are over-extrapolated.** A hedging register in the prompt becomes a hedging register in the output, amplified, because the model treats the surface as the target rather than as evidence. U2’s apology is the example.
3. **Stable habits are not reproduced.** Regularities the user applies across every instance of a genre, which never appear in any single prompt because the user assumes them, are invisible. U1’s crediting of collaborators is the example.

Seepage also explains a common subjective experience: a generic assistant feels briefly personalised, because it does echo something real about the user, and then fails on exactly the parts of a task the user did not spell out. The echo is real; the model beneath it is absent.

3.3. Why accumulating preferences treats the symptom

The remedy that deployed systems offer for seepage is to accumulate preference notes from past conversations and reinsert them. This is a reasonable engineering response to the observation that the assistant forgets, and it does address forgetting. It does not address seepage, because the notes are observed at the level of preferences while the cause of variation across users sits one layer below.

Two properties of accumulated notes make this concrete. First, they are *topically bound*: a note distilled from a conversation about choosing a laptop is about laptops, and generalises to an unrelated writing task only by accident. Section 7 gives an empirical form of this observation, in that the preference-notes condition was not distinguishable from chance on forced-choice identification of the user’s own voice. Second, they are *uninterpretable against each other*: a flat list has no structure that would let a system notice that two notes conflict, or that one has gone stale, or that a third is a surface expression of the same underlying disposition as the first two. Structure is what makes those operations definable, which is the argument for a schema and the subject of the next section.

4. The Atomic User Model

4.1. The metaphor, and what it commits us to

AUM represents a person as a Nucleus surrounded by four shells, on the metaphor of an atom with a stable centre and reactive outer layers (Figure 3). The metaphor is doing work rather than decorating, and it commits the representation to two properties that a flat schema cannot express.

Differential stability. Identity is stable while behaviour adapts. The Nucleus changes on the timescale of years, the inner shells on months, the outer shells on weeks. A representation that treats every field as equally durable will either refresh core values as often as it refreshes a communication habit, which is wasteful and invasive, or refresh a communication habit as rarely as it refreshes core values, which makes it wrong.

Differential observability. A stranger sees the outermost shell, a colleague sees more, a close friend or family member sees much more, and the Nucleus is visible only to the self. This layering echoes the onion model of social penetration theory (Altman and Taylor, 1973; Carpenter and Greene, 2015), and it has a direct computational consequence: the fields most useful for deep personalisation are exactly the fields most costly to expose. A representation that does not encode the gradient cannot reason about the trade-off.

Both properties bear on Section 5. The first motivates differential update rates as a field attribute. The second is why a budgeted payload is a privacy mechanism and not only an efficiency one.

4.2. Formal definition

The AUM of a user u is a set of typed fields

$$A_u = \{f_1, f_2, \dots, f_n\}, \quad (1)$$

where each field f_i is a tuple

$$f_i = \langle \text{shell}(f_i), \text{name}(f_i), \text{value}(f_i), \tau(f_i) \rangle. \quad (2)$$

Here $\text{shell}(f_i) \in \mathcal{S}$ with

$$\mathcal{S} = \{\text{Nucleus}, \text{Shell 1}, \text{Shell 2}, \text{Shell 3}, \text{Shell 4}, \text{CrossShell}\}, \quad (3)$$

$\text{name}(f_i)$ is a fixed slot name drawn from the specification, $\text{value}(f_i)$ is a short natural-language string, and $\tau(f_i)$ is a nominal update period. Three properties of this definition are load-bearing.

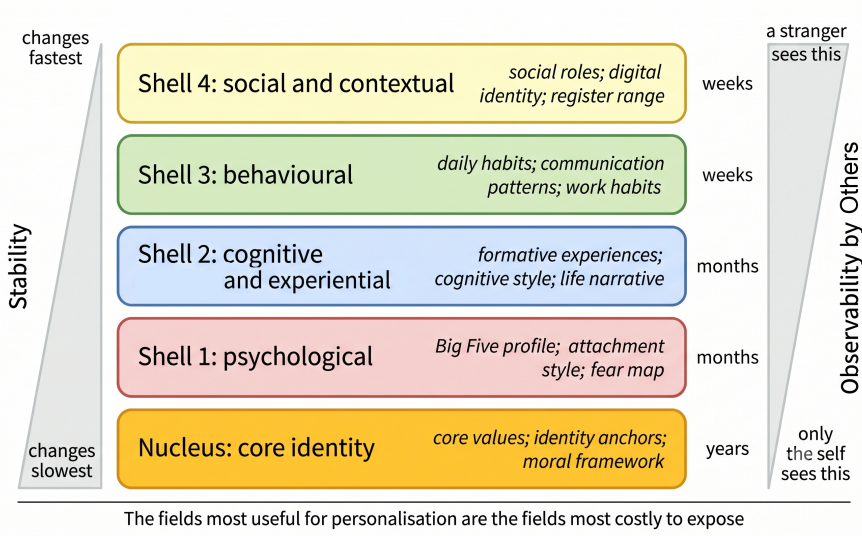


Figure 2: Differential stability across tiers. The nominal update period τ falls from years at the Nucleus to weeks at Shell 4, while observability by others rises in the opposite direction. A schema that does not encode this gradient cannot express a refresh policy, and cannot reason about the cost of exposing a field.

1. **Values are natural language, not vectors.** A field reads “I credit whoever reviewed a draft before I ask them for anything else”, not a coordinate. This is what makes the model scrutable in the sense of [Kay and Kummerfeld \(2012\)](#): the user can read, edit or delete any field without an interpretation layer.
2. **Slots are declared in advance.** The set of (shell, name) pairs is fixed by the specification, so two users’ models are structurally comparable and a selection function over shells is definable. This is the property that distinguishes AUM from bottom-up accumulations of free-form propositions ([Shaikh et al., 2025](#)).
3. **Update rate is a field attribute.** τ ranges over years for Nucleus fields, months for Shell 1 and Shell 2, and weeks for Shell 3 and Shell 4. We do not specify an update mechanism in this paper and return to the omission in Section 9.5.

4.3. The five tiers

Nucleus: core identity.. The Nucleus holds what changes most slowly and what, if changed, makes the person a different agent: core values, fundamental beliefs about self, others and life, identity anchors of the form “I am the kind of person who ...”, a moral framework, existential orientation, and life purpose. The framing follows Erikson’s account of identity ([Erikson, 1968](#)). Operationally, a Nucleus field is one whose violation the user would describe as being untrue to themselves rather than as an inconsistency.

Shell 1: psychological core.. Shell 1 explains how the Nucleus is expressed. It holds personality structure, incorporating but not reducing to inventories such as the Five-Factor Model ([McCrae and John, 1992](#); [Soto and John, 2017](#)); emotional architecture; the attachment system ([Bowlby, 1969](#)); motivational drivers; a fear map; and stress response. Trait inventories, including Myers-Briggs type where a user knows it, are fields inside Shell 1, not substitutes for the model. Given the psychometric objections to type-based instruments ([Pittenger, 2005](#); [Stein and Swan, 2019](#)),

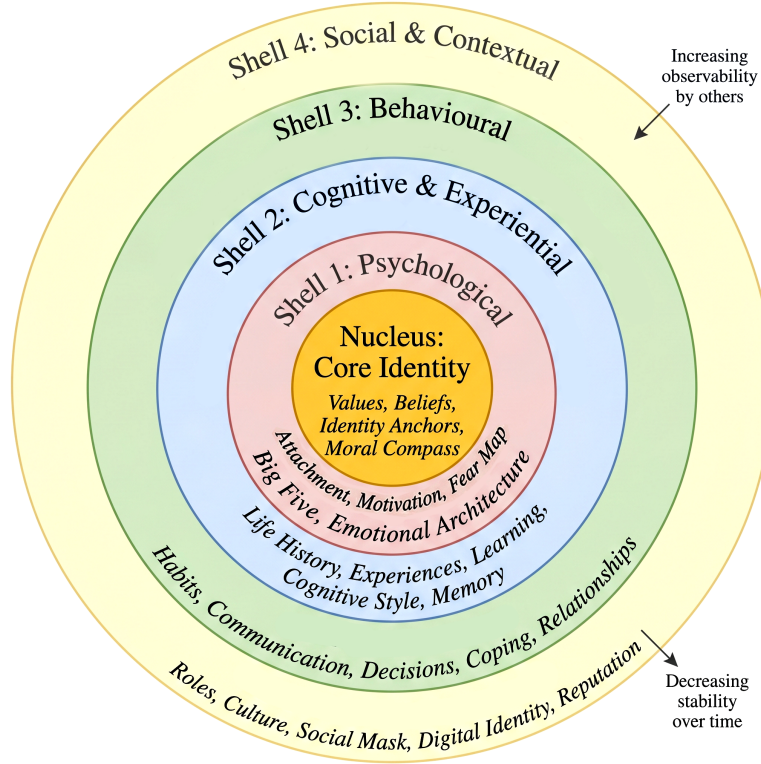


Figure 3: The Atomic User Model. A Nucleus of core identity is surrounded by four shells. Example fields per shell are shown in italics. Stability decreases and observability by others increases from the centre outwards.

a schema that demotes an inventory to one field among thirty-two is a feature rather than a compromise.

Shell 2: cognitive and experiential.. Shell 2 is life history as it bears on cognition: formative experiences, a record of difficult episodes, learning style, cognitive style, the personal narrative the user tells about their own life, and knowledge base. It is the tier that explains why a given Shell 1 profile produces the particular Shell 3 behaviours it does in this particular person.

Shell 3: behavioural patterns.. Shell 3 holds daily habits, communication patterns, decision behaviour, coping mechanisms, relationship behaviour and work habits. This is the tier most often mistaken for personality in deployed systems, because it is the tier that observation reaches. It is the output of the inner shells, not their cause, which is why a system that models only Shell 3 can predict what a user did and not what they would do in a situation it has not seen.

Shell 4: social and contextual.. Shell 4 is the interface to the world: social roles, cultural conditioning, digital identity, reputation, the social mask a person presents (Goffman, 1959), and register range. Most LLM personalisation today operates on the outer half of Shell 4, because that is the part chat history exposes.

Cross-shell integration.. Three entries sit across tiers rather than inside one. *Internal conflicts* record cases where a Shell 3 behaviour contradicts a Nucleus value, which is the machinery that lets a system notice that a stored preference is out of character rather than merely old. *Growth*

Table 2: The reduced 32-field instrument used in both studies. Six fields per tier plus two cross-shell entries. Update period τ is the nominal refresh interval for the tier.

Tier	τ	Fields
Nucleus	years	Core Values; Fundamental Beliefs; Identity Anchors; Moral Framework; Existential Orientation; Life Purpose
Shell 1 psychological	months	Big Five Profile; Emotional Architecture; Attachment Style; Motivational Drivers; Fear Map; Stress Response
Shell 2 cognitive	months	Formative Experiences; Difficult Episodes; Learning Style; Cognitive Style; Life Narrative; Knowledge Base
Shell 3 behavioural	weeks	Daily Habits; Communication Patterns; Decision Behaviour; Coping Mechanisms; Relationship Behaviour; Work Habits
Shell 4 social	weeks	Social Roles; Cultural Conditioning; Digital Identity; Reputation; Social Mask; Register Range
CrossShell	months	Internal Conflicts; Authenticity Index

trajectory records the direction of deliberate change. The *authenticity index* measures alignment between what the inner shells hold and what the outer shells present, and is the field a system would consult before, for example, matching a user’s professional register in a private message.

4.4. The full specification and the reduced instrument

The full specification holds approximately sixty fields across the five tiers plus cross-shell entries; Appendix [Appendix A](#) lists it. The studies in this paper use a reduced 32-field instrument, six fields per tier plus two cross-shell entries, shown in Table 2. The reduction was made for a practical reason: a simulated participant completing sixty fields in character produces noticeably thinner answers per field than one completing thirty-two, and we preferred fewer, denser fields. Section 8 examines what happens to retrieval as the instrument grows towards the full size.

4.5. Two design commitments

Local residence.. AUM contains inner-shell content of a kind that has no analogue in a preference list. It should reside on the user’s device or in a user-controlled vault, never in a third-party log. We state this as an architectural commitment rather than a policy preference because the retrieval design of Section 5 makes it enforceable: only the selected payload need ever leave the device, and Section 7 measures how small that payload can be without loss. Contextual integrity ([Nissenbaum, 2011](#)) gives the frame for what such a boundary is for.

User authorship.. Every field is a human-inspectable string that the user can read, edit or delete. This places AUM in the scrutable-personalisation tradition ([Kay and Kummerfeld, 2012](#); [Czarkowski and Kay, 2002](#); [Bakalov et al., 2010](#)) and answers the interpretability objection to trait-vector user models ([Ji et al., 2023](#); [Kang et al., 2023](#)). A budgeted payload strengthens the commitment: eight short strings are few enough to show the user before they are sent, where thirty-two are not.

5. Personality-Aware Retrieval

AUM is a representation of a person; it requires a pipeline to be useful. This section specifies one, in which AUM is treated as an index over a person and queried under a context budget rather than injected wholesale.

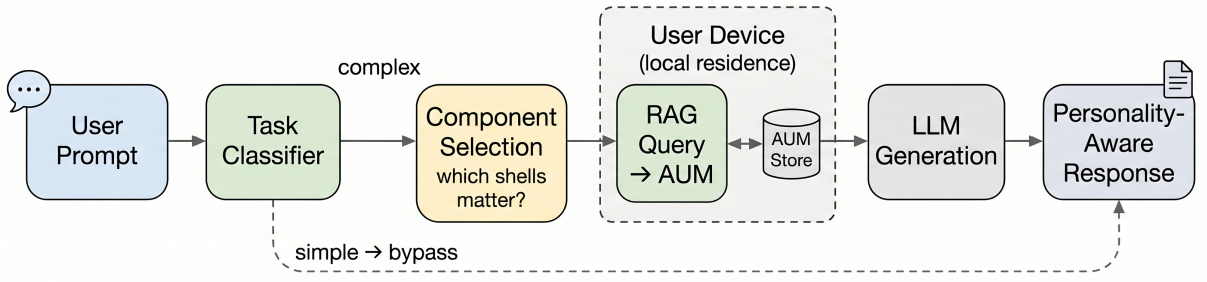


Figure 4: The personality-aware retrieval pipeline. A task classifier routes simple requests around the user model entirely. Complex requests pass through component selection, which restricts the search to the shells the task type touches, and then through a budgeted retrieval that returns k fields. The AUM store remains on the user’s device; only the payload leaves it.

5.1. The four stages

Given a prompt q from user u , the pipeline of Figure 4 proceeds as follows.

(i) *Task identification*.. A classifier assigns q a task type t and a complexity label. A request is *simple* if a correct answer is the same regardless of who asked, such as a unit conversion or a factual lookup, and *complex* if the form of an acceptable answer depends on who is asking. Simple requests bypass the user model, which matters both for latency and for disclosure: a system that consults an intimate user model to convert kilometres into miles is leaking for no benefit. Section 7.7 reports how accurately this stage runs.

(ii) *Component selection*.. A selection function $\sigma : \mathcal{T} \rightarrow 2^{\mathcal{S}}$ restricts the search space to the shells that a task type touches. The mapping used throughout is given in Table 3; CrossShell is always admissible. The admissible pool is

$$\text{pool}(t) = \{i : \text{shell}(f_i) \in \sigma(t) \cup \{\text{CrossShell}\}\}. \quad (4)$$

If $|\text{pool}(t)| < k + 4$ the pool is topped up with the most query-relevant fields from outside $\sigma(t)$. Without this top-up a large k would silently collapse to “every field $\sigma(t)$ admits”, which would make the $k = 16$ condition a different experiment from the $k = 4$ and $k = 8$ conditions rather than a third point on one curve.

(iii) *Retrieval*.. A retriever selects a payload $P(q, u, k) \subseteq A_u$ with $|P| = k \ll n$ by maximising the objective of Section 5.2 over $\text{pool}(t)$.

(iv) *Generation*.. The model is conditioned on (q, P) . The payload is inserted as a delimited block ahead of the user’s request; the exact template is in Appendix B.

5.2. The payload objective

Let $\mathbf{f}_i$ denote an embedding of field i and $\mathbf{q}$ an embedding of the query, both ℓ_2 -normalised. The obvious objective is mean relevance,

$$F_{\text{rel}}(S) = \frac{1}{|S|} \sum_{i \in S} \cos(\mathbf{f}_i, \mathbf{q}), \quad (5)$$

Table 3: The component-selection function $\sigma(t)$. CrossShell is admissible for every complex task type. Factual requests are routed around the user model entirely, which is what the empty set denotes.

Task type t	$\sigma(t)$
communication	Shell 3, Shell 4
self-presentation	Shell 4, Nucleus, Shell 1
relational-evaluative	Shell 1, Shell 3, Nucleus
planning-decision	Shell 3, Shell 1, Shell 2
explanatory	Shell 2, Shell 3
creative	Shell 1, Shell 2
factual (simple)	$\emptyset$ (bypass)

but this is separable: it decomposes into per-item scores, so its exact maximiser over a pool is obtained by sorting, and no search is required or useful. Under a context budget, mean relevance is also the wrong objective. Eight near-duplicate fields waste seven slots, and a budgeted payload should therefore trade relevance against diversity in the manner of maximal marginal relevance (Carbonell and Goldstein, 1998) and of result diversification more broadly (Clarke et al., 2008; Agrawal et al., 2009). We use

$$F(S) = \underbrace{\frac{1}{|S|} \sum_{i \in S} \cos(\mathbf{f}_i, \mathbf{q})}_{\text{relevance}} - \lambda \underbrace{\frac{1}{|S|(|S| - 1)} \sum_{\substack{i, j \in S \\ i \neq j}} \cos(\mathbf{f}_i, \mathbf{f}_j)}_{\text{redundancy}}, \quad (6)$$

with $\lambda \geq 0$ the redundancy weight, and the payload is

$$P(q, u, k) = \arg \max_{\substack{S \subseteq \text{pool}(t) \\ |S|=k}} F(S). \quad (7)$$

At $\lambda = 0$, Equation 6 reduces to Equation 5 and the problem is a sort. For $\lambda > 0$ the redundancy term couples the items and the problem becomes combinatorial, with $\binom{|\text{pool}(t)|}{k}$ candidate payloads and no closed form. We use $\lambda = 0.15$ throughout, and Section 8 reports what changes as λ varies.

Two remarks on the structure of the objective. First, because it uses a mean rather than a sum and subtracts a mean pairwise similarity, it is not monotone submodular in general, so the classical greedy guarantee of Nemhauser et al. (1978) does not transfer; how close greedy comes is an empirical question, answered in Section 8. Second, the redundancy term is defined over the payload only, not over the user’s history, which is what distinguishes it from novelty in a search result list.

5.3. Retrievers

Four retrievers are implemented, plus exhaustive search wherever the search space is small enough to enumerate.

Dense top-k. Take the k highest-relevance fields, ignoring redundancy. This is the retrieval-augmented generation default and is the exact maximiser of Equation 5. Cost: one objective evaluation.

Forward greedy.. Start from $S = \emptyset$ and repeatedly add the field that most increases F . Cost: $O(k \cdot |\text{pool}|)$ objective evaluations. Strong, cheap and deterministic; it is the baseline that matters.

Random k .. Draw k fields uniformly from $\text{pool}(t)$. This is a control rather than a method. Without it, an improvement for a retrieved payload is equally consistent with the hypothesis that any k short strings about the user help, which is not the claim under test.

Artificial fish swarm.. A population of p fish, each a k -subset of the pool, evolves for T iterations under three behaviours drawn from Li et al. (2002). Algorithm 1 gives the procedure. In *prey*, a fish attempts up to *try* random local swaps and accepts the first improvement, falling back to a random walk. In *swarm*, a fish moves toward the centre of its visual neighbourhood if that centre is better and the neighbourhood is not crowded. In *follow*, it moves toward the best neighbour under the same conditions. Distance between fish is $d(a, b) = 1 - |a \cap b|/k$, and a move toward a target replaces the lowest-relevance member of the current subset with a member of the target. The global best is retained, so the returned payload never degrades across iterations. In the simulation study we use $p = 12$, $T = 8$, $\text{visual} = 6$, $\text{try} = 4$ and crowd factor 0.75; Section 8 shows that this fixed budget is the source of the algorithm’s weakness rather than the behaviours are.

5.4. Why local residence becomes measurable

The design objective is to maximise style fidelity subject to $|P| \leq k$. That constraint is what separates AUM from a prompt prefix, and it is where the commitments of Section 4 stop being rhetorical. If k fields suffice, the remaining $n - k$ never leave the device, and the disclosure per query is bounded by construction rather than by policy. The size of the achievable k is therefore not only an efficiency number but a privacy number, and Section 7 measures it.

6. Simulation Study: Design

6.1. Why simulate, and what a simulation can establish

Participants in this study are language-model-simulated, not human. We state this first, and repeat it in the abstract, because it governs how every number in Section 7 should be read.

The method has a precedent and a critique, and both are relevant. Argyle et al. (2023) showed that conditioning a language model on demographic backstories reproduces aggregate patterns found in human survey data, and Park et al. (2024) showed that grounding agents in structured self-report improves individual-level simulation over demographic conditioning alone, which is the closest precedent for what we do here. Against this, Bisbee et al. (2024) document instability and misestimation in synthetic survey responses, and Wang et al. (2025) show that replacing human participants can misportray and flatten identity groups.

We therefore fix what the study is for. A simulation of this kind *can* establish that a pipeline runs end to end; that a budgeted retrieval over a structured user model recovers fields that a downstream generator uses; that the resulting text differs measurably from text produced without the model; and that this holds under controls which rule out the most obvious alternative explanations. It *cannot* establish that a real person recognises their own voice, because the entity doing the recognising here is a judge model reading a persona description, which is a considerably easier task. It also cannot speak to how a person would feel about a system holding such a model, which is a question for the human study described in Section 9.6.

Algorithm 1 Artificial fish swarm over k -subsets

Require: pool $\mathcal{P}$, budget k , objective F , population p , iterations T , visual v , tries try , crowd δ

Ensure: payload S^* with $|S^*| = k$

```
1: Pop  $\leftarrow \{S_1, \dots, S_p\}$ , each a uniform random  $k$ -subset of  $\mathcal{P}$ 
2:  $S^* \leftarrow \arg \max_{S \in \text{Pop}} F(S)$ 
3: for  $\iota = 1$  to  $T$  do
4:   for all  $S \in \text{Pop}$  do
5:      $C \leftarrow \{\text{PREY}(S)\}$ 
6:      $N \leftarrow \{S' \in \text{Pop} : S' \neq S, 1 - |S \cap S'|/k \leq v/k\}$ 
7:     if  $N \neq \emptyset$  and  $|N|/p \leq \delta$  then
8:        $S_c \leftarrow$  the  $k$  most frequent members across  $N$  ▷ swarm centre
9:       if  $F(S_c) > F(S)$  then  $C \leftarrow C \cup \{\text{MOVE TOWARD}(S, S_c)\}$ 
10:      end if
11:       $S_b \leftarrow \arg \max_{S' \in N} F(S')$ 
12:      if  $F(S_b) > F(S)$  then  $C \leftarrow C \cup \{\text{MOVE TOWARD}(S, S_b)\}$ 
13:      end if
14:    end if
15:     $S \leftarrow \arg \max_{S' \in C \cup \{S\}} F(S')$ 
16:  end for
17:   $S^* \leftarrow \arg \max\{F(S^*), \max_{S \in \text{Pop}} F(S)\}$  ▷ elitism
18: end for
19: return  $S^*$ 

20: function PREY( $S$ )
21:   for  $r = 1$  to  $try$  do
22:      $S' \leftarrow S$  with one or two members replaced at random from  $\mathcal{P} \setminus S$ 
23:     if  $F(S') > F(S)$  then return  $S'$ 
24:   end if
25: end for
26:   return  $S$  with one member replaced at random ▷ random walk
27: end function

28: function MOVE TOWARD( $S, S_t$ )
29:   replace  $\arg \min_{i \in S} \cos(\mathbf{f}_i, \mathbf{q})$  with a uniformly chosen member of  $S_t \setminus S$ 
30:   return the result
31: end function
```

There is a second and less obvious reason to simulate first. The human study this substitutes for would require participants to disclose inner-shell content about themselves: fears, difficult episodes, moral commitments. Running that study before knowing whether the architecture works at all would spend real disclosure on a question that a simulation can answer. We regard establishing feasibility in simulation as the ethically prior step, not merely the cheaper one.

Table 4: The sixteen simulated participants. Big Five z -scores are shown as O/C/E/A/N. Idiolect markers are abbreviated; the full declarations are in Appendix [Appendix C](#).

#	Label	O/C/E/A/N	Idiolect markers (abridged)
0	meticulous planner	+0.2/+1.8/-0.6/+0.4/+0.9	numbered points, explicit dates, ask in first sentence
1	warm over-explainer	+0.6/+0.1/+1.1/+1.7/+0.5	long pleasantries, hedges, pre-emptive apology, “we” for “I”
2	blunt minimalist	+0.3/+0.8/-0.9/-1.4/-0.7	lowercase, fragments, no greeting or sign-off, two lines
3	associative creative	+1.9/-0.8/+0.7/+0.6/+0.3	starts mid-thought, parentheses, concrete images, ends early
4	analytical flat	+0.9/+1.3/-1.2/-0.5/-0.9	conclusion first, quantifies, “approximately”, no exclamations
5	bright energetic	+1.0/+0.2/+1.9/+1.2/-0.4	exclamations, short punchy sentences, direct address
6	dry sceptic	+1.1/+0.6/-0.5/-1.1/+0.8	understatement, undercutting asides, rhetorical questions
7	formal correct	-0.3/+1.5/-0.2/+0.9/+0.2	full salutations, passives, “I should be grateful if”, no contractions
8	scattered fast	+1.4/-1.6/+1.3/+0.3/+1.1	mid-paragraph topic changes, follow-up corrections, abbreviations
9	nurturing relational	+0.4/+0.3/+0.5/+1.9/+0.7	names the other’s feeling first, softens every imperative
10	competitive driver	+0.5/+1.4/+1.2/-1.3/-0.5	power verbs, deadlines as commitments, reframes asks as opportunities
11	reflective introvert	+1.2/+0.9/-1.8/+0.8/+0.6	short paragraphs with pauses, precise word choice, no small talk
12	storytelling extrovert	+1.3/-0.2/+1.7/+1.0/-0.2	opens with a scene, second person, builds to the ask, quotes dialogue
13	pragmatic efficient	-0.4/+1.6/+0.1/+0.2/-1.0	imperatives, concrete nouns, one line per item
14	quirky metaphorical	+1.8/+0.4/+0.3/+0.7/+0.4	cross-domain comparisons, semicolons, deliberate archaisms
15	stoic calm	+0.1/+1.0/-0.7/+0.5/-1.8	even sentence length, no intensifiers, facts then action

6.2. Simulated participants

Sixteen personas are declared in code rather than sampled by a model, so that a run is reproducible from a seed and the population is auditable. Each persona carries Big Five z -scores in $[-2, 2]$, a one-line life context, a formative-experience seed on which Shell 2 can draw, and concrete idiolect markers, which are what make style measurable rather than merely asserted. The z -scores were spread deliberately: no two personas share a sign pattern across all five traits, and every trait has both tails represented, so that style differences are not confounded with a single dominant axis. Table 4 lists the set. Appendix [Appendix C](#) gives the full declarations.

6.3. Instrument population

Each persona completes the 32-field instrument of Table 2 in character, by guided self report: the agent is instructed to fill every field with one short first-person sentence of roughly ten to

Table 5: The task battery. The six complex tasks form the evaluation set. The four simple tasks are used only to score the classifier’s routing decision and never enter the main results.

ID	Task	Type
T1	Requesting a deadline extension from a colleague	communication
T2	Declining an invitation from a colleague	communication
T3	Writing a short public post about your work	self-presentation
T4	Giving written feedback on a junior collaborator’s draft	relational-evaluative
T5	Planning a two-day trip	planning-decision
T6	Explaining a technical result to a friend outside the field	explanatory
S1 to S4	Four factual lookups (boiling point, unit conversion, date, arithmetic)	factual

twenty-five words, specific and concrete rather than generic. Responses are parsed against the declared schema and rejected if any shell is missing or any field is empty, with up to six retries and an explicit corrective instruction on each. The exact prompt is in Appendix [Appendix B](#).

Populating a user model by self report inherits a known limitation: self and other observers know different things about a person, and the self is not the better source for every trait ([Vazire, 2010](#); [Connelly and Ones, 2010](#)). For a simulated participant whose ground truth is a declared persona this is less severe than it would be for a human, since the agent has access to its own declaration, but the resulting instrument is still a self-description rather than an observation. Section [9.5](#) returns to what this means for a human deployment.

6.4. Task battery, prompt surfaces, and references

Six style-sensitive tasks form the evaluation battery, chosen to span the task types of Table [3](#): requesting a deadline extension (communication), declining an invitation from a colleague (communication), writing a short public post about one’s work (self-presentation), giving written feedback on a junior collaborator’s draft (relational-evaluative), planning a two-day trip (planning-decision), and explaining a technical result to a friend outside the field (explanatory). Four additional simple tasks, all factual lookups, exist only to exercise and score the classifier’s simple-versus-complex branch and are excluded from the main results. Table [5](#) lists the battery.

For each complex task an agent produces two texts, in an order that is the single most important correctness property of the design.

1. The **prompt surface**: the message this person would type to an assistant, in their own voice, at their own level of detail and formality, capped at sixty words. This is the personality-seepage surface of Section [3](#), and it is the *only* user-derived text the retriever ever sees.
2. The **reference**: what this person would write themselves, with no assistant involved, matching their usual length, structure and habits. This is the ground truth against which every condition is scored, and it is generated without sight of the prompt surface.

Passing the reference to the retriever would give every AUM-R condition an oracle that the Full AUM and preference-notes baselines never receive, and would make the comparison meaningless. The separation is enforced in code.

Table 6: Conditions. The main block appears in Table 8; the control block isolates one component at a time and appears in Table 10. Only the four conditions marked ID are shown to the identification judge.

Condition	Block	Injected beyond the prompt
Generic ID	main	Nothing
Preference notes ID	main	Flat bullets distilled from four held-out prior conversations
Full AUM ID	main	All 32 fields as JSON
AUM-R, $k = 4$	main	$\sigma(t)$ + swarm retrieval, four fields
AUM-R, $k = 8$ ID	main	$\sigma(t)$ + swarm retrieval, eight fields
AUM-R, $k = 16$	main	$\sigma(t)$ + swarm retrieval, sixteen fields
Random $k = 8$	control	Eight fields drawn uniformly from the same admissible pool
No $\sigma(t)$, $k = 8$	control	Swarm retrieval over all shells, no component selection
Dense top- k , $k = 8$	control	Relevance-only top eight, redundancy term removed
Greedy, $k = 8$	control	Forward greedy on the same objective

6.5. The preference-notes baseline

The baseline that matters is not the generic assistant but the flat preference notes that deployed memory features actually store, so it has to be built the way such features build it. Each agent generates four short transcripts of past assistant conversations, four to six turns each, on topics drawn from a held-out list that shares no item with the evaluation battery: choosing between two laptops, a recurring appliance problem, a week of simple meals, what to do with a free Saturday, picking a book, and replying to an awkward group message. A separate summariser, with no access to the persona and no access to the AUM, then reads those transcripts and writes at most twelve flat preference bullets.

Two decisions here are deliberate. The topics are held out so that the baseline cannot contain a rehearsal of the evaluation tasks. The summariser is blind to the persona and the user model so that the baseline is a lossy distillation of observed conversation, which is what a deployed system has, rather than a partial copy of the treatment, which is what concatenating Shell 3 and Shell 4 fields would produce.

6.6. Conditions

Ten conditions, all sharing a base model and a prompt, split into a main block and a control block (Table 6).

Each control removes exactly one component, so that a null result localises. Random k separates “the right fields” from “any k fields about the user”. No $\sigma(t)$ asks whether component selection contributes beyond retrieval. Dense top- k asks whether the redundancy term earns its place. Greedy asks whether the swarm earns its cost.

6.7. Measures

Fidelity. A judge model rates, on a five-point scale, the statement that a candidate text reads like something the person would have written. The judge sees the persona and one text the person genuinely wrote, and never sees the AUM. Withholding the user model from the judge is not

incidental: showing it would present the judge with the same fields the AUM-R payload contained, allowing it to reward surface agreement with the payload rather than similarity to the person.

Identification.. A forced choice among the four ID conditions, presented in an order shuffled per trial, with chance at 25%. The judge again sees the persona and the reference. Position bias in LLM judges is documented (Wang et al., 2024a), so option order is randomised per trial and tested post hoc in Section 7.7.

Style similarity.. The cosine between LUAR authorship embeddings (Rivera-Soto et al., 2021) of the candidate and the participant’s own reference, computed locally. Authorship representations are used rather than semantic embeddings for a specific reason: a semantic embedding scores any two deadline-extension emails as highly similar because they are about the same thing, and the property we need is sensitivity to *who* wrote a text and insensitivity to *what* it is about (Wegmann et al., 2022; Wang et al., 2023). The pipeline raises rather than falling back if the authorship model cannot be loaded, because a silent substitution would change what the column means while leaving its heading intact. A semantic cosine is computed in addition and reported separately, never as style similarity.

Context.. The number of tokens in exactly the string prepended to the prompt, counted with the generator’s own tokeniser. This is a measurement, not a word count, so that the ratio reported in Section 7 is exact.

Judge independence.. The judge model differs from the generator by construction, and the pipeline refuses to run otherwise. LLM evaluators favour their own generations (Panickssery et al., 2024), so a run in which judge and generator coincide produces a fidelity column that mostly measures self-preference.

6.8. Models, seeds and cost

Table 7 records the configuration. Every API call is cached on disk under a hash of its arguments, so a re-run after a crash or a code change costs nothing for work already completed, and the run is reproducible from the manifest plus a seed.

6.9. Analysis plan

The unit of analysis is the (participant, task) cell averaged over seeds, giving $16 \times 6 = 96$ paired observations per condition. Seeds are repeated measures of the same cell rather than additional participants, so averaging before testing is what keeps the degrees of freedom honest; treating three seeds $\times$ 96 cells as 288 independent observations would inflate every test.

For each planned contrast we report the mean paired difference, a 95% bootstrap confidence interval over 10,000 resamples, a paired t statistic, a Wilcoxon signed-rank statistic, Cohen’s d_z and Cliff’s delta. The Wilcoxon test is primary for fidelity, because a five-point rating scale is ordinal; the paired t is primary for the continuous style measure. The Holm procedure corrects within metric. Identification uses an exact binomial test against 25% chance with a Wilson interval.

Contrasts were fixed before the run and are those listed in Section 7. Supplementary analyses reported in Section 7.7 were specified after inspecting the main results and are labelled as such; we do not present them as confirmatory.

Table 7: Reproducibility record for the simulation study. Token and cost figures are read from the run ledger, not estimated.

Item	Value
Generator model (also plays the participant)	<code>gpt-5.6-luna</code>
Judge model (fidelity and identification)	<code>gpt-5.6-terra</code>
Field and query embeddings	<code>text-embedding-3-small</code>
Authorship-style embeddings	<code>rrivera1849/LUAR-MUD</code> , run locally on CPU
Generation temperature	0.7
Persona temperature	0.8
Judge temperature	0.0
Participants $\times$ tasks $\times$ seeds	$16 \times 6 \times 3$
Seeds	0, 1, 2
Conditions	10
Instrument size n	32 fields
Redundancy weight λ	0.15
Swarm population, iterations	12, 8
Total API calls	13,281
Input tokens	3,869,382
Output tokens	760,598
Total cost	USD 4.81

7. Simulation Study: Results

7.1. Main comparison

Table 8 gives the main block, and Table 10 the planned contrasts. Three findings carry the section.

Finding 1: a budgeted payload matches the whole model at under a quarter of the context.. AUM-R at $k = 8$ was not distinguishable from Full AUM on fidelity ($\Delta = +0.02$, 95% CI $[-0.07, +0.11]$, $p = 1$ after Holm correction) while injecting 211 tokens against 915, that is 23.0%. Figure 5 plots the whole condition set against injected context on a logarithmic axis; the point of the figure is the vertical gap between preference notes and AUM-R at comparable cost, and the horizontal gap between AUM-R and Full AUM at comparable fidelity.

Finding 2: structure beats flat notes, by a moderate effect.. AUM-R at $k = 8$ exceeded preference notes on fidelity by $+0.24$ (95% CI $[+0.14, +0.33]$, Wilcoxon $W = 376.5$, $p = 7.2 \times 10^{-5}$ after Holm correction, $d_z = +0.50$, Cliff’s $\delta = 0.35$, $n = 96$ cells), and exceeded the generic baseline by $+0.27$ ($d_z = +0.57$). On LUAR style similarity the same ordering holds with a smaller effect: $+0.013$ over preference notes ($p = 0.006$) and $+0.024$ over generic ($p < 10^{-5}$, $d_z = +0.53$). Full AUM also beat preference notes ($+0.22$, $p = 4 \times 10^{-4}$), which is the expected ordering: the gain comes from having a structured model at all, and the retrieval stage recovers it cheaply rather than adding to it.

Table 8: Main results. Fidelity and identification are judge ratings on behalf of simulated participants; style is LUAR authorship cosine against the participant’s own reference; semantic is a general-purpose embedding cosine, reported for comparison only and never as style; context is tokens injected beyond the prompt. Means over $16 \times 6 = 96$ cells and three seeds, with 95% bootstrap intervals on fidelity.

Condition	Fidelity (1 to 5)	95% CI	Ident. (%)	Style (LUAR)	Context (tokens)
Generic (prompt only)	2.65	[2.45, 2.84]	6.9	0.674	0
Preference notes	2.68	[2.48, 2.88]	14.9	0.684	164
Full AUM (all 32 fields)	2.90	[2.71, 3.08]	35.4	0.705	915
AUM-R, $k = 4$	2.86	[2.68, 3.06]	n/a	0.701	111
AUM-R, $k = 8$ (ours)	2.91	[2.74, 3.09]	42.7	0.698	211
AUM-R, $k = 16$	2.98	[2.78, 3.17]	n/a	0.701	406
<i>Controls, all at $k = 8$</i>					
Random k fields	2.89	[2.69, 3.07]	n/a	0.697	211
No $\sigma(t)$	2.95	[2.75, 3.14]	n/a	0.703	209
Dense top- k	2.92	[2.74, 3.10]	n/a	0.699	210
Greedy	2.95	[2.75, 3.14]	n/a	0.700	210

Table 9: Forced-choice identification. Chance is 25%; $n = 288$ presentations per condition ($96 \text{ cells} \times 3 \text{ seeds}$). Intervals are Wilson; p is an exact binomial test against chance in the upper tail.

Condition	Hits	%	95% CI	p vs 25%
AUM-R, $k = 8$ (ours)	123	42.7	[37.1, 48.5]	4.2×10^{-11}
Full AUM (all fields)	102	35.4	[30.1, 41.1]	5.3×10^{-5}
Preference notes	43	14.9	[11.3, 19.5]	1.00
Generic (prompt only)	20	6.9	[4.5, 10.5]	1.00

Finding 3: identification separates the conditions sharply.. Forced-choice identification of the participant’s own voice reached 42.7% for AUM-R at $k = 8$ (95% CI [37.1, 48.5], $p = 4.2 \times 10^{-11}$ against 25% chance) and 35.4% for Full AUM ([30.1, 41.1], $p = 5.3 \times 10^{-5}$). Preference notes reached 14.9% and generic 6.9%, neither above chance (Table 9, Figure 7). That both user-model conditions sit above chance while both baselines sit below it is the cleanest separation in the study, and it is the result most directly relevant to the inversion argument of Section 3: notes distilled from unrelated past conversations carry topic, not person.

7.2. The budget curve

Fidelity rose monotonically across the budgets tested: 2.86 at $k = 4$, 2.91 at $k = 8$, 2.98 at $k = 16$ (Figure 6). Neither adjacent step was individually significant after correction ($k = 16$ against $k = 8$: $\Delta = +0.06$, $p = 0.74$; $k = 8$ against $k = 4$: $\Delta = +0.05$, $p = 1$), so the curve is best read as a gradual improvement rather than a threshold. Two things follow. First, $k = 8$ is not a saturation point, and a system with a looser budget should use more fields. Second, and more usefully for the privacy argument, there is no evidence of degradation as further fields enter the payload: the concern that irrelevant fields would dilute a payload and hurt output quality is not

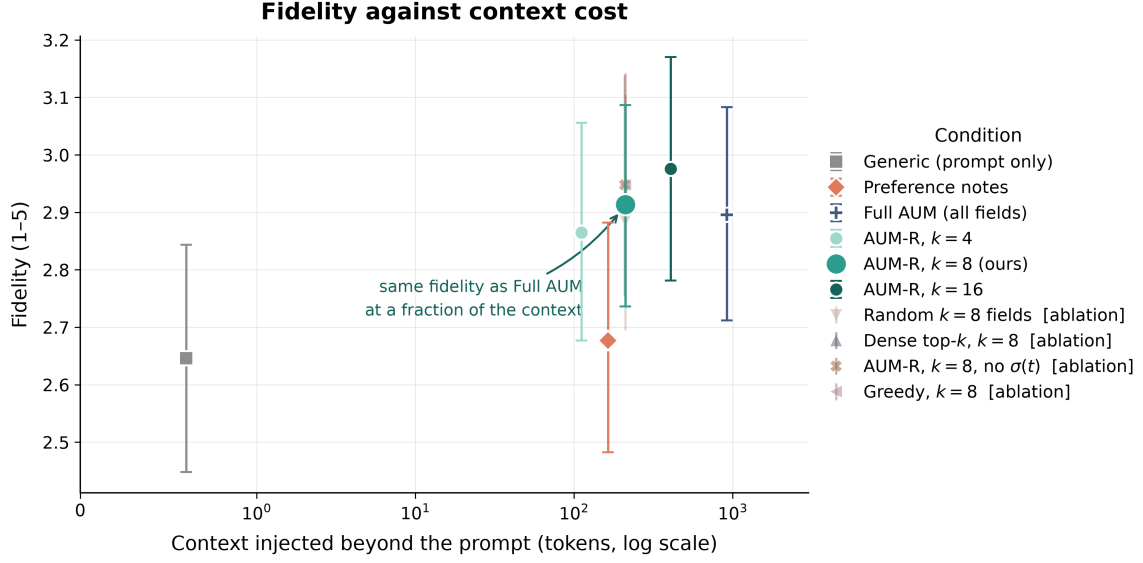


Figure 5: Fidelity against injected context on a logarithmic axis. Eight retrieved fields reach the fidelity of the entire 32-field user model at 23% of its context. Error bars are 95% bootstrap intervals over the 96 (participant, task) cells.

borne out at these budgets, so the argument for a small k rests on cost and disclosure rather than on quality.

7.3. Style similarity

Figure 8 shows the distribution of LUAR authorship cosine by condition, and Figure 9 the general-purpose semantic cosine for the same trials. The two behave differently in an instructive way. The authorship metric separates the conditions in the same order as fidelity, with a compressed range (0.674 to 0.705 across all ten conditions). The semantic metric separates them far less (0.593 to 0.620), which is what one expects given that every candidate for a given trial is an attempt at the same task: semantic similarity is dominated by topic, and topic is held constant by design. We report the semantic column only to make that contrast visible, and we would regard a paper that reported it as “style similarity” as having measured the wrong thing.

7.4. Four controls return null

All four controls returned null, on both metrics, and we take each in turn because each bounds the claim differently.

The right k fields, or merely k fields? AUM-R at $k = 8$ was not distinguishable from eight fields drawn uniformly at random from the same admissible pool ($\Delta = +0.028$, $p = 1$). This is the most consequential null in the study. It says that once component selection has narrowed the pool to roughly seventeen fields, which eight of them are injected did not measurably matter. The gain in Findings 1 to 3 is therefore attributable to injecting structured, task-plausible fields about the person, not to ranking within that set.

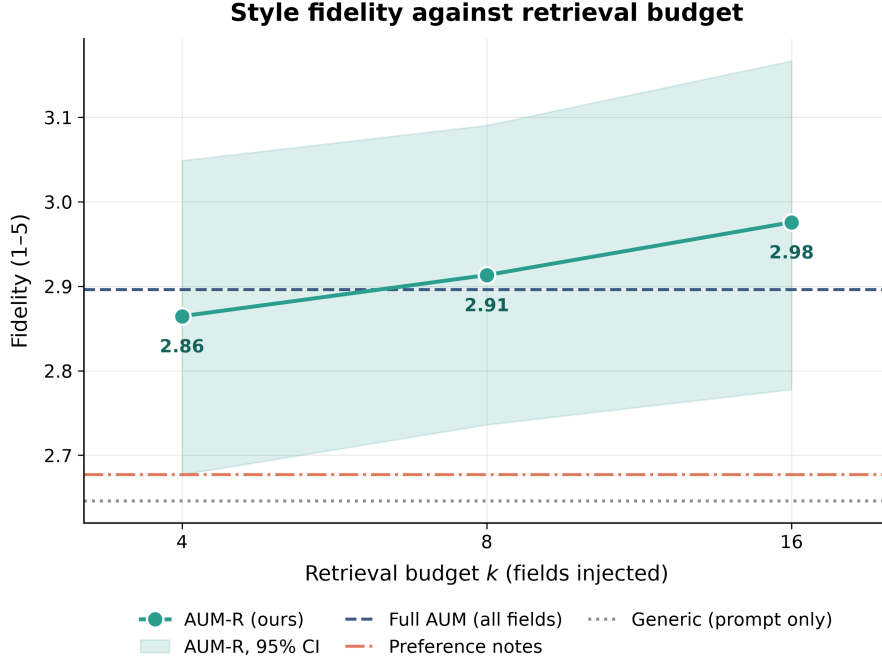


Figure 6: Fidelity against retrieval budget k , with the Full AUM and preference-notes levels marked. Fidelity is still rising at the largest budget tested, so $k = 8$ is a trade rather than a saturation point.

Does $\sigma(t)$ contribute beyond retrieval? Removing component selection and retrieving over all thirty-two fields did not hurt; it was very slightly better and not significantly so ($\Delta = -0.035$, $p = 0.98$). Given the preceding null this is coherent rather than surprising: if ranking within the pool does not matter, restricting the pool cannot matter either, so long as the restriction does not exclude something essential. Section 7.5 shows that $\sigma(t)$ nonetheless selects sensibly, which is a weaker but still useful property.

Does the redundancy term earn its place? A relevance-only top- k was indistinguishable from the full objective ($\Delta = -0.007$, $p = 1$). Section 8 shows this is a property of the operating point rather than of the objective: at $\lambda = 0.15$ and a pool of seventeen, the redundancy penalty rarely changes which eight fields win.

Does the swarm earn its cost? It did not. AUM-R did not beat forward greedy on generated style ($\Delta = -0.035$, $p = 1$), and it did not reach a higher value of its own objective. Table 11 gives the retrieval diagnostics: the swarm reached $F = 0.224$ using 1,736 objective evaluations, where greedy reached $F = 0.226$ using 108 and dense top- k reached $F = 0.225$ using one. Random reached $F = 0.183$, which confirms that the objective is meaningful even though optimising it harder did not help downstream. Figure 10 shows the quality and cost side by side.

What the four nulls jointly establish.. They localise the effect. Something about injecting a structured, task-plausible slice of a user model changes generated style substantially; nothing about which slice, or about how hard one searches for it, changes it further at this scale. That is a narrower claim than the one we set out to test, and it is the claim the data supports. Section 8 takes

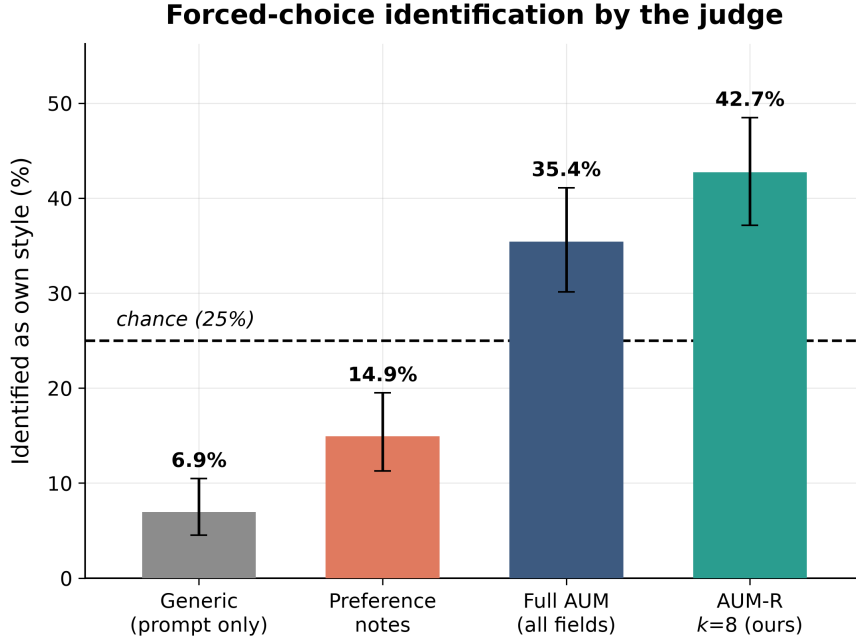


Figure 7: Forced-choice identification by the judge, with Wilson intervals and the 25% chance line. Both user-model conditions sit above chance; both baselines sit below it.

the optimisation question out of the language-model setting entirely and asks where, if anywhere, the search does start to matter.

7.5. Which shells each task type draws on

Although component selection did not change downstream quality, the payloads it produced are interpretable, and inspecting them is a check that the pipeline is doing what the specification says. Figure 11 shows the share of retrieved payload drawn from each shell, broken down by task type, and Table D.23 gives the same figures numerically.

The pattern matches the mapping in Table 3 without having been forced to. Communication tasks drew 44.4% of their payload from Shell 3 and 42.3% from Shell 4. Self-presentation drew most heavily on the Nucleus (37.2%) and Shell 4 (41.2%), which is what one would expect of a task in which a person decides what to project. Planning and decision tasks drew 49.5% from Shell 3 and split the rest between Shell 1 and Shell 2. Relational-evaluative tasks, which involve judging another person’s work, drew 24.7% from the Nucleus, the second-highest Nucleus share in the study, alongside 45.6% from Shell 3. Explanatory tasks drew 30.0% from Shell 2, which holds knowledge base and cognitive style, the highest Shell 2 share of any task type. Cross-shell fields appeared in every task type at between 7.6% and 14.8%.

This is a weak result and we present it as one. It shows that the representation is legible and that retrieval over it behaves sensibly, not that legibility improves output. It matters mainly for the scrutability argument: a user inspecting why the assistant produced a particular draft can be shown a payload whose composition is defensible in ordinary language.

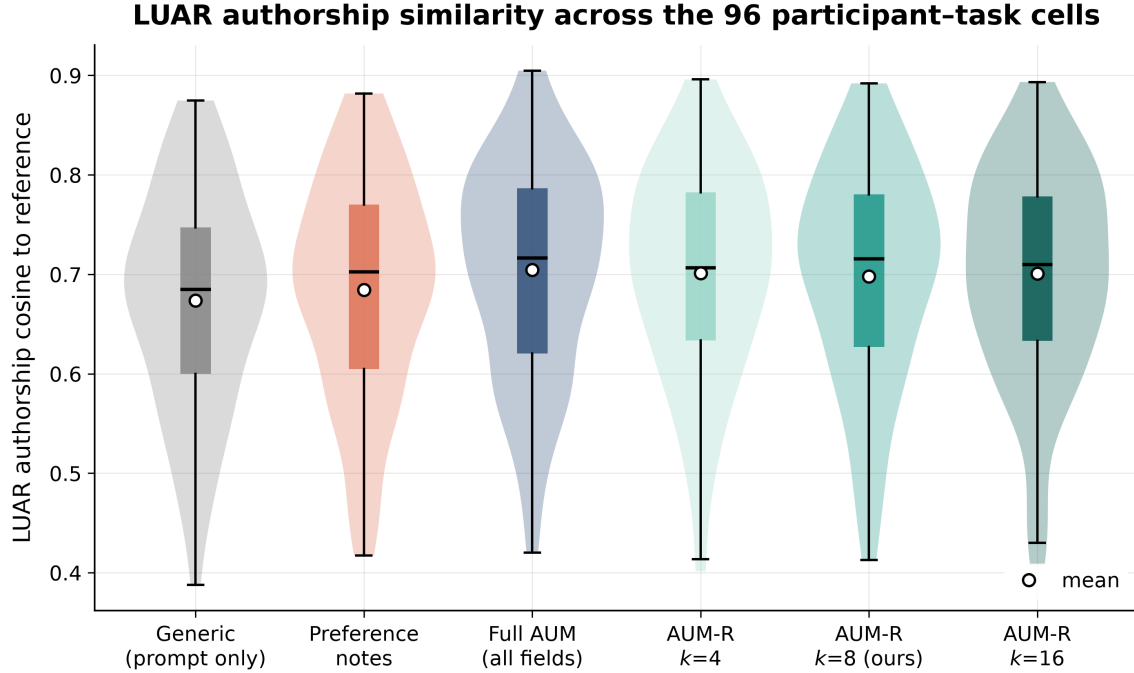


Figure 8: Distribution of LUAR authorship-embedding cosine between each generated text and the participant’s own reference, by condition.

7.6. Which tasks the effect holds for

All six tasks favoured AUM-R in the mean (Table 12, Figure 12), but the sizes differ in a way worth naming. The largest effects are on tasks with the loosest genre conventions: explaining something to a friend outside the field (+0.375) and declining an invitation (+0.312). The smallest, and effectively zero, is the deadline-extension email (+0.021, favouring in only 18.8% of participants). That is the most conventionalised task in the battery: a deadline-extension email has a widely shared expected form, and a generic assistant already produces it. Where the genre supplies the structure, the user model has little left to contribute; where it does not, the model fills the gap. This is the task-level counterpart of the participant-level pattern in Section 7.8.

7.7. Robustness

Four supplementary analyses address the objections a reader is most likely to raise. All were specified after the main results and are reported as exploratory.

The classifier stage works.. Section 5 posits a routing stage, and an unevaluated first stage is a weak point in any pipeline. Over the full battery the classifier assigned the correct complexity label in 98.8% of trials (83 of 84, Wilson interval [93.6, 99.8]), with all twelve simple trials routed to bypass and 71 of 72 complex trials routed to the user model. Six-way task-type accuracy on complex trials was 90.3% (65 of 72, [81.3, 95.2]) against 16.7% chance. Table 13 gives per-type recall and Figure 13 the confusion matrix. The single systematic error is that explanatory requests are read as communication in a third of cases, which is intelligible: explaining a result to a friend is a communication act, and the two $\sigma(t)$ sets share Shell 3.

Semantic similarity across the 96 participant-task cells (secondary measure)

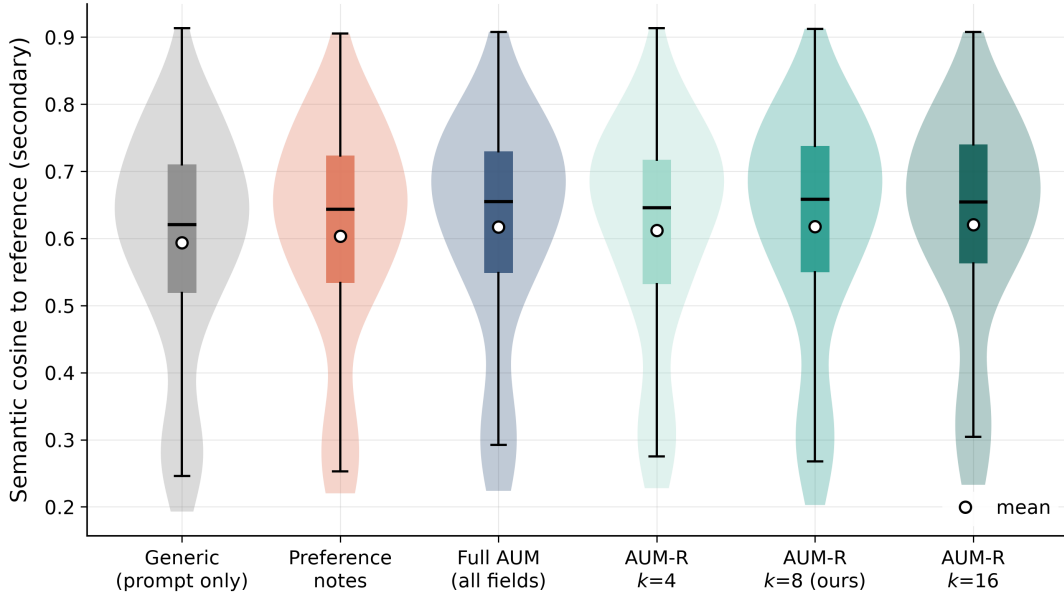


Figure 9: The same trials under a general-purpose semantic embedding. The compression relative to Figure 8 is the point: a semantic metric mostly measures that every candidate addresses the same task, which is the wrong invariance for a style claim.

The effect is not a length artefact.. A judge asked whether a text reads like a person’s own writing might be rewarded largely by length agreement. It is: pooled across all trials, the absolute log ratio of candidate length to reference length correlates strongly and negatively with fidelity (Spearman $\rho = -0.511$, $p = 3 \times 10^{-191}$, $n = 2880$). Matching the reference’s length is a large part of reading like the person, which is unsurprising and not itself a confound.

The question is whether the treatment effect survives adjustment for it. It does. Regressing the paired fidelity difference (AUM-R $k = 8$ minus preference notes) on the paired difference in absolute log length ratio leaves an intercept of $+0.227$ (95% CI $[+0.127, +0.327]$, $p = 1.9 \times 10^{-5}$) against a raw difference of $+0.236$, and the covariate slope is not significant (-0.155 , $p = 0.519$). In other words the two conditions did not differ appreciably in length agreement, and the effect is not carried by it. Table 14 gives the per-condition figures. The generic condition is the outlier: it produced the shortest outputs (96 words against a 142-word reference mean), the worst length agreement (0.007 on the absolute log ratio scale relative to the next condition) and the highest rate of degenerate outputs (7.6% of generic trials produced fewer than fifteen words, typically a clarifying question rather than an attempt at the task).

The effect is broad, not driven by outliers.. Twelve of sixteen participants and six of six tasks favoured AUM-R in the mean. Of the 96 cells, 51.0% favoured AUM-R, 33.3% were exact ties on the ordinal scale, and 15.6% favoured preference notes. Figure 15 gives a per-participant forest plot, and Table 15 the numbers. Two participants show small negative effects, the meticulous planner (-0.111) and the reflective introvert (-0.056), and neither interval excludes zero.

The judge is usable but not sharp, and is not position-biased.. Across the three seeds the fidelity judge achieved $\text{ICC}(2,1) = 0.57$ over 960 targets, which by the conventional bands is moderate

Table 10: Planned contrasts, paired over the 96 (participant, task) cells. Δ is the mean paired difference; the interval is a 95% bootstrap interval over 10,000 resamples; p is Holm-corrected within metric, with Wilcoxon primary for fidelity and paired t primary for style.

Metric	Contrast	Δ	95% CI	d_z	p
fidelity	AUM-R $k=8$ vs Preference notes	+0.236	[+0.142, +0.330]	+0.50	0.0001
fidelity	AUM-R $k=8$ vs Generic	+0.267	[+0.177, +0.365]	+0.57	0.0001
fidelity	Full AUM vs Preference notes	+0.219	[+0.115, +0.323]	+0.42	0.0004
fidelity	AUM-R $k=8$ vs Full AUM	+0.017	[−0.069, +0.108]	+0.04	1.0000
fidelity	AUM-R $k=16$ vs AUM-R $k=8$	+0.063	[−0.031, +0.156]	+0.13	0.7436
fidelity	AUM-R $k=8$ vs AUM-R $k=4$	+0.049	[−0.031, +0.125]	+0.12	1.0000
fidelity	AUM-R $k=8$ vs Random $k=8$	+0.028	[−0.052, +0.108]	+0.07	1.0000
fidelity	AUM-R $k=8$ vs no $\sigma(t)$	−0.035	[−0.111, +0.045]	−0.09	0.9813
fidelity	AUM-R $k=8$ vs Dense top- k	−0.007	[−0.083, +0.073]	−0.02	1.0000
fidelity	AUM-R $k=8$ vs Greedy	−0.035	[−0.111, +0.042]	−0.09	1.0000
style	AUM-R $k=8$ vs Preference notes	+0.013	[+0.006, +0.021]	+0.36	0.0057
style	AUM-R $k=8$ vs Generic	+0.024	[+0.015, +0.033]	+0.53	0.0000
style	Full AUM vs Preference notes	+0.020	[+0.012, +0.029]	+0.47	0.0001
style	AUM-R $k=8$ vs Full AUM	−0.007	[−0.014, +0.001]	−0.18	0.5889
style	AUM-R $k=8$ vs Random $k=8$	+0.001	[−0.005, +0.007]	+0.04	1.0000
style	AUM-R $k=8$ vs no $\sigma(t)$	−0.006	[−0.012, +0.001]	−0.17	0.6055
style	AUM-R $k=8$ vs Dense top- k	−0.001	[−0.007, +0.005]	−0.03	1.0000
style	AUM-R $k=8$ vs Greedy	−0.002	[−0.009, +0.005]	−0.06	1.0000

reliability (Shrout and Fleiss, 1979; Koo and Li, 2016). We report the intraclass correlation rather than a chance-corrected agreement coefficient because the ratings are ordinal and the three seeds are exchangeable rather than fixed coders, which is the condition under which the coefficient families diverge (Krippendorff, 2004; Artstein and Poesio, 2008). Exact agreement between seed pairs ranged from 42.0% to 43.2% and agreement within one scale point from 85.6% to 87.4%, with Spearman correlations of 0.55 to 0.60. The judge used the whole scale rather than collapsing to the middle: 13.8% of ratings were 1, 25.2% were 2, 25.8% were 3, 30.9% were 4 and 4.3% were 5. On identification, choices did not cluster significantly by option slot despite the per-trial shuffling ($\chi^2 = 7.37$, 3 d.f., $p = 0.061$), with slot shares of 22.9%, 29.2%, 27.4% and 20.5%. That p is close enough to the conventional threshold that we would not describe position bias as absent, only as not detected at this sample size. Figure 16 shows both diagnostics.

7.8. Personalisation buys most where the default voice is furthest away

The per-participant breakdown suggested a pattern that we then tested directly. Fidelity in the generic condition is a direct measure of how well the un-personalised assistant already writes like a given participant: it is the score the default voice earns without any user model at all. If personalisation is doing what it claims, its benefit should be largest where that baseline is lowest.

It is. Across the sixteen participants, generic-condition fidelity correlates negatively with the treatment effect (Spearman $\rho = -0.61$, $p = 0.013$; Pearson $r = -0.60$, $p = 0.013$), and the relationship holds at cell level with more power and a smaller coefficient ($\rho = -0.27$, $p = 0.007$, $n = 96$). Splitting participants at the median generic fidelity, those the default serves worst gain +0.333 (95% CI [+0.181, +0.486]) and those it serves best gain +0.139 ([+0.035, +0.243]); the

Table 11: Retrieval diagnostics at $k = 8$, averaged over all trials. $F(S)$ is the value of Equation 6 reached; evaluations counts calls to F ; the pool is the size of $\text{pool}(t)$ after component selection. The swarm spent sixteen times greedy’s budget to reach a marginally lower objective value.

Retriever	$F(S)$	Evaluations	Seconds	Pool
Greedy, $k = 8$	0.2256	108.5	0.0014	16.9
Dense top- k , $k = 8$	0.2246	1.0	0.0001	16.9
AUM-R (swarm), $k = 8$	0.2240	1,735.7	0.0196	16.9
Random, $k = 8$	0.1833	1.0	0.0001	16.9
AUM-R, $k = 4$	0.2512	1,748.5	0.0166	16.9
AUM-R, $k = 16$	0.2030	1,746.8	0.0213	20.0
AUM-R, $k = 8$, no $\sigma(t)$	0.2376	1,791.6	0.0193	32.0

Table 12: Effect of AUM-R at $k = 8$ over preference notes, by task. Δ is the mean paired difference in fidelity over the sixteen participants; “favouring” is the share of participants for whom the difference is positive. p is an uncorrected Wilcoxon signed-rank test within task.

ID	Task	Δ	95% CI	Favouring	p
T6	Explain to a friend	+0.375	[+0.042, +0.708]	56.3%	0.045
T2	Decline invitation	+0.312	[+0.062, +0.604]	50.0%	0.061
T5	Trip planning	+0.250	[+0.083, +0.396]	62.5%	0.024
T3	Public post	+0.229	[+0.063, +0.396]	68.8%	0.018
T4	Feedback on draft	+0.229	[0.000, +0.479]	50.0%	0.116
T1	Deadline extension	+0.021	[−0.063, +0.104]	18.8%	0.783

difference between the two groups is in the expected direction but does not itself reach significance (Mann-Whitney $p = 0.096$). Figure 17 plots the relationship.

The extremes are interpretable. The blunt minimalist, whose lowercase fragments and absent sign-offs are about as far from a default assistant register as the battery contains, scores 1.78 under generic and gains +0.611, the largest effect in the study. The meticulous planner, whose numbered points and explicit dates are close to what an assistant produces unprompted, scores 2.94 under generic and gains nothing (−0.111). The default voice already was that person.

One caveat is important and we state it rather than burying it. The same correlation computed on the LUAR style metric rather than on judge fidelity is in the same direction but does not reach significance ($\rho = -0.12$, $p = 0.242$). The pattern is therefore established in the judge’s ratings and not, on this sample, in the automatic measure. We report it because its design implication is substantial and because suppressing a result that did not replicate across both metrics would be worse than reporting it with its caveat attached.

8. Scaling Study of the Retriever

8.1. Separating two questions

Section 7.4 reported that the swarm retriever neither improved generated style over forward greedy nor reached a higher value of its own objective, and that it spent roughly sixteen times as

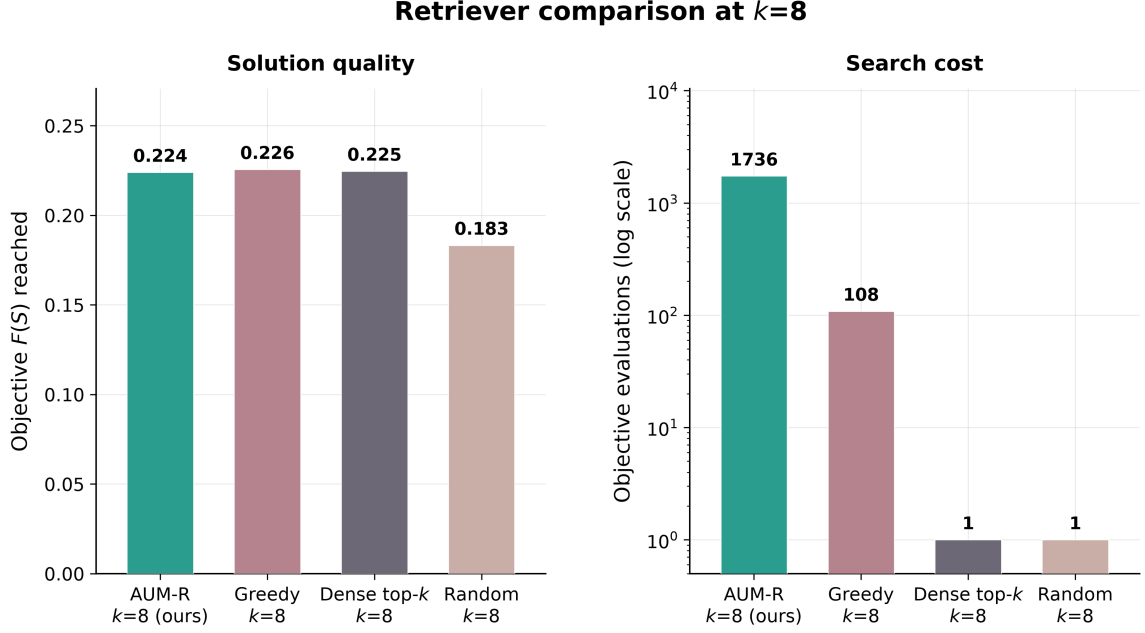


Figure 10: Retriever comparison at $k = 8$: objective value reached and objective evaluations spent. The swarm is both slightly worse and sixteen times more expensive than forward greedy at this scale.

many objective evaluations doing so. That is a result about one operating point: a reduced 32-field instrument, a pool of about seventeen admissible fields after component selection, and $k = 8$. It is not a result about payload selection in general, and treating it as one would be as much of an error as the reverse.

This section separates the two questions. It removes the language model entirely and studies the optimisation problem on its own, sweeping the instrument size n , the context budget k and the redundancy weight λ . The question it answers is narrow and answerable: for which (n, k, λ) does a population method reach a better value of Equation 6 than forward greedy, and at what cost. No API call is involved, so the whole sweep is free, deterministic from a seed, and can be re-run by a reader.

8.2. Instance model

Field embeddings from real text are not isotropic Gaussian vectors. Fields belonging to the same shell talk about related things and therefore cluster, and that clustering is precisely what makes the redundancy term non-trivial. Each synthetic instance is generated with an explicit cluster structure. For n fields distributed over m shells in d dimensions,

$$\mathbf{c}_s \sim \text{Uniform}(\mathcal{S}^{d-1}), \quad s = 1, \dots, m, \quad (8)$$

$$\mathbf{f}_i = \text{normalise}\left(\rho \mathbf{c}_{s(i)} + \sqrt{1 - \rho^2} \boldsymbol{\varepsilon}_i\right), \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I), \quad (9)$$

$$\mathbf{q} = \text{normalise}\left(\sum_{s \in \mathcal{A}} w_s \mathbf{c}_s + \sigma_q \boldsymbol{\varepsilon}_q\right), \quad (10)$$

where $s(i)$ is the shell of field i , $\mathcal{A}$ is a random subset of shells of size two, and $w_s \sim \text{Uniform}(0.5, 1.5)$. The query construction mimics a task that draws on two shells, which is what Table 3 specifies. The

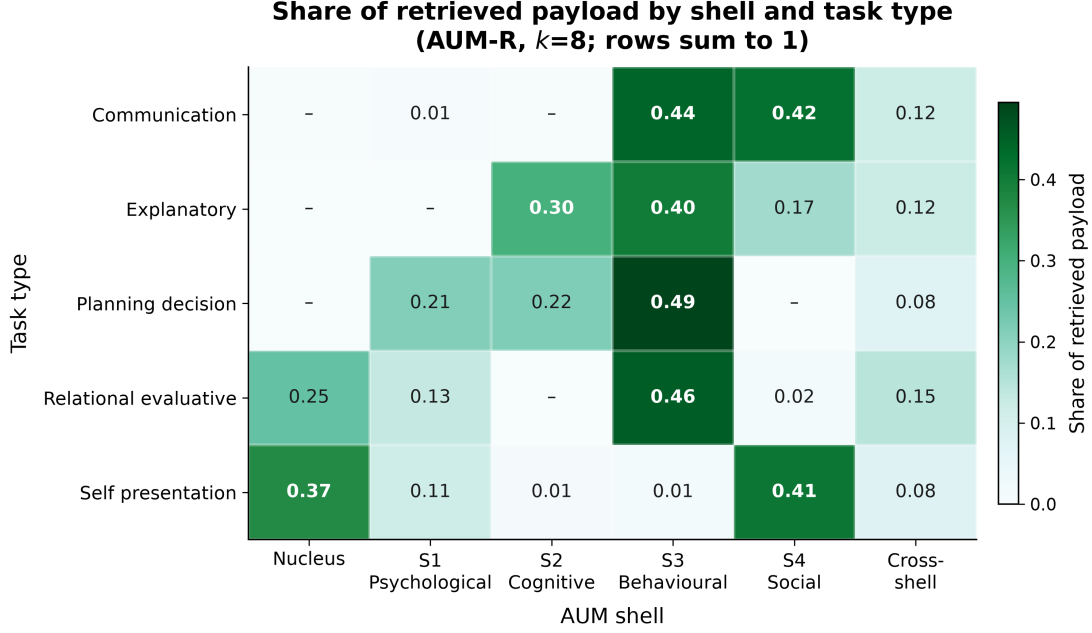


Figure 11: Share of the retrieved payload drawn from each shell, by task type. Rows sum to 100%. The composition tracks Table 3 without having been constrained to.

parameter ρ controls how tightly a shell clusters; $\rho = 0$ recovers the isotropic case and makes the redundancy term nearly inert. We use $\rho = 0.65$, $d = 64$, $m = \max(2, \lfloor n/6 \rfloor)$ and 25 independent instances per cell.

8.3. Methods and budgets

Six retrievers are compared, plus exhaustive search wherever $\binom{n}{k} \leq 60,000$.

Dense top- k the relevance-only maximiser of Equation 5, one objective evaluation.

Greedy forward greedy on the full objective.

Random k fields drawn uniformly, the floor.

AFSA (study budget) the swarm with exactly the population and iteration counts used in Section 6, $p = 12$ and $T = 8$, held fixed as n grows. This is what the simulation actually ran.

AFSA (scaled budget) population and iterations grown with the search space: $p = \text{clip}(\lceil 2 \log \binom{n}{k} \rceil, 12, 60)$ and $T = \text{clip}(\lceil 1.2 \log \binom{n}{k} \rceil, 8, 40)$. The constants are set so that $(n, k) = (32, 8)$ reproduces roughly the study’s own budget and everything larger receives proportionally more.

AFSA (greedy warm start) the scaled-budget swarm with one member of the initial population replaced by the greedy solution. This is the obvious hybrid and the fairest test of whether a population method can add anything on top of greedy.

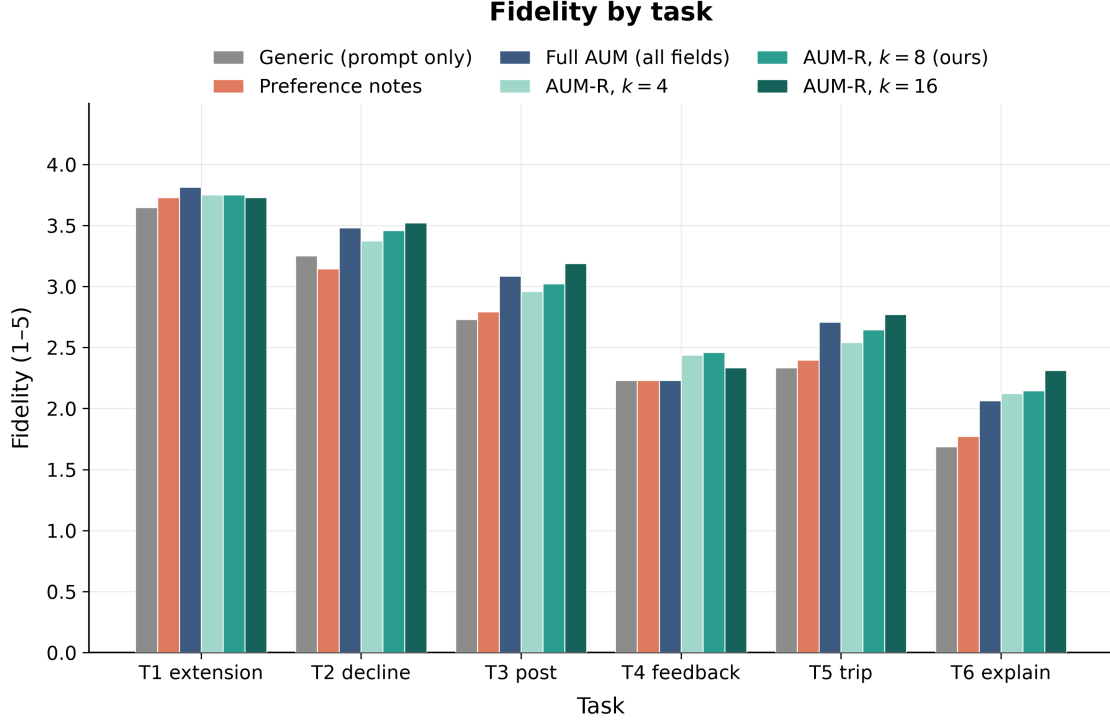


Figure 12: Fidelity by task and condition. The deadline-extension task, the most conventionalised in the battery, shows the smallest spread across conditions.

The sweep covers $n \in \{16, 32, 64, 128, 256\}$, $k \in \{4, 8, 16, 32\}$ with $k \leq n/2$, and $\lambda \in \{0, 0.15, 0.30, 0.60\}$, giving 68 cells and 1,700 instances. Quality is reported as the relative gap to the best solution found for that instance by any method, which equals the exhaustive optimum wherever that is available.

8.4. Results

Forward greedy is very hard to beat.. Across the entire sweep, greedy stayed within 0.10% of the best solution found (Table 16, Figure 18). Where exhaustive search was tractable, it returned the exact optimum in 77.7% of instances. It is not a straw man, and the fact that Equation 6 is not submodular in general (Nemhauser et al., 1978) does not appear to cost it much in practice on instances with this cluster structure.

The study's swarm budget, not the swarm, is what failed.. AFSA at the fixed budget of Section 6 degraded from a 0.44% gap at $n = 16$ to 43.23% at $n = 256$, and found the exact optimum in only 46.3% of the tractable instances against greedy's 77.7%. It is a poor optimiser at any n beyond the smallest, and it was already worse than greedy at the $n = 32$ operating point the simulation used. This is the direct explanation of the null in Section 7.4: the swarm did not lose because population methods are unsuited to this problem, it lost because it was given a budget that does not grow with the search space.

Growing the budget helps, up to a point.. The scaled-budget swarm stayed within 0.6% up to $n = 64$ and found the exact optimum in 84.7% of tractable instances, exceeding greedy's 77.7%. Beyond

Table 13: Task classifier performance, pooled over three seeds. Recall is within gold task type on complex trials only.

Gold type	n	Recall	Dominant confusion
planning-decision	12	100.0%	n/a
relational-evaluative	12	100.0%	n/a
communication	24	95.8%	planning-decision (1)
self-presentation	12	83.3%	creative (2)
explanatory	12	66.7%	communication (4)
Overall, six-way		90.3%	chance 16.7%
Simple vs complex		98.8%	chance 50%

Table 14: Output length by condition, against a reference mean of 141.6 words. “Agreement” is the mean absolute log ratio of candidate to reference length, where lower is closer. “Stub” is the share of trials producing fewer than fifteen words.

Condition	Words	95% CI	Agreement	Stub
Generic (prompt only)	96.3	[71.4, 124.9]	1.007	7.6%
Preference notes	112.7	[85.6, 143.6]	0.920	5.9%
Full AUM	114.5	[88.5, 143.6]	0.844	3.5%
AUM-R, $k = 4$	112.2	[86.3, 140.9]	0.881	5.2%
AUM-R, $k = 8$ (ours)	114.0	[86.9, 145.0]	0.862	4.5%
AUM-R, $k = 16$	114.1	[86.6, 145.3]	0.840	2.4%
Greedy, $k = 8$	109.7	[83.6, 139.3]	0.859	4.9%
Dense top- k , $k = 8$	109.3	[83.8, 138.2]	0.860	4.2%

that it degrades again (5.46% at $n = 128$, 12.59% at $n = 256$), because even a budget growing with $\log \binom{n}{k}$ is a vanishing fraction of the search space.

Warm-starting from greedy dominates.. The warm-started swarm stayed within 0.05% at every scale tested and found the exact optimum in 96.3% of tractable instances. Against greedy head to head it won or tied in 100% of instances (16.0% strict wins, 84.0% ties, mean $\Delta = +0.001$), which is what one expects of a method that starts from greedy’s answer and keeps the better of the two. It costs roughly seven times greedy’s evaluations (Table 17).

The redundancy weight decides whether any of this matters.. Figure 21 and Table 18 isolate λ at $n = 64$, $k = 8$. At $\lambda = 0$ the objective is separable, dense top- k is exactly optimal by construction, and every search method is wasted effort. As λ rises the relevance-only solution degrades: a 0.32% gap at $\lambda = 0.15$, 1.51% at $\lambda = 0.30$ and 7.68% at $\lambda = 0.60$, at which point greedy also begins to lose ground (1.73%). This locates the regime in which combinatorial search over a payload is worth performing at all: it is governed by how much a system penalises near-duplicate fields, and at the $\lambda = 0.15$ we used, the penalty is mild enough that dense top- k is nearly optimal. That is the second half of the explanation for the null against dense top- k in Section 7.4.

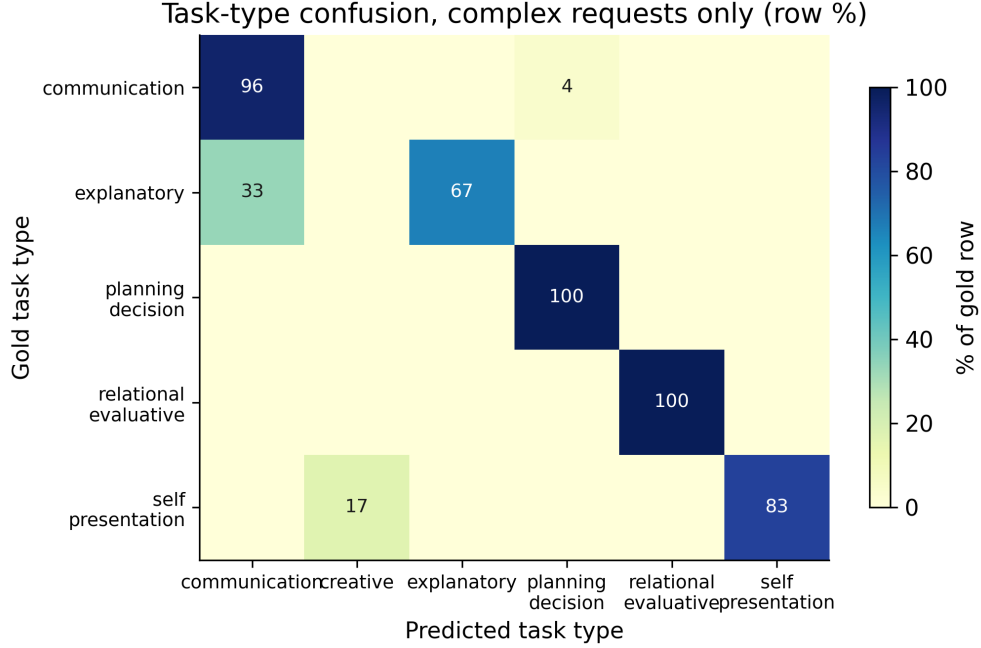


Figure 13: Task-type confusion for complex requests, as row percentages. The one systematic error, explanatory read as communication, reflects a genuine overlap between the two categories.

8.5. What this means for deployment

Three recommendations follow, and we state them as recommendations because they change what we would build.

1. **Use forward greedy.** At any instrument size we tested it is within 0.1% of the best solution found, at a cost of $O(kn)$ evaluations. For an on-device retriever over sixty fields this is a few hundred dot products and is not worth improving on.
2. **If a population method is used, warm-start it from greedy.** A cold-started swarm is strictly worse than greedy at every scale beyond the smallest. A warm-started one never loses, reaches the exhaustive optimum in 96% of tractable instances, and costs about seven times greedy. Whether that is worth paying depends on whether the objective value is what one actually cares about, which Section 7.4 suggests it may not be.
3. **Set λ deliberately, or drop the redundancy term.** Below roughly $\lambda = 0.3$ the redundancy penalty rarely changes which fields win, and a system that is not going to tune λ should use dense top- k and save the search entirely. Above it, the penalty matters and greedy is the right tool.

The broader methodological point is that the null in Section 7.4 was not uninformative. Reading it required taking the optimisation question out of the language-model setting, where a downstream metric with a moderate effect size cannot resolve a 0.2% difference in objective value, and asking it where it could be answered.

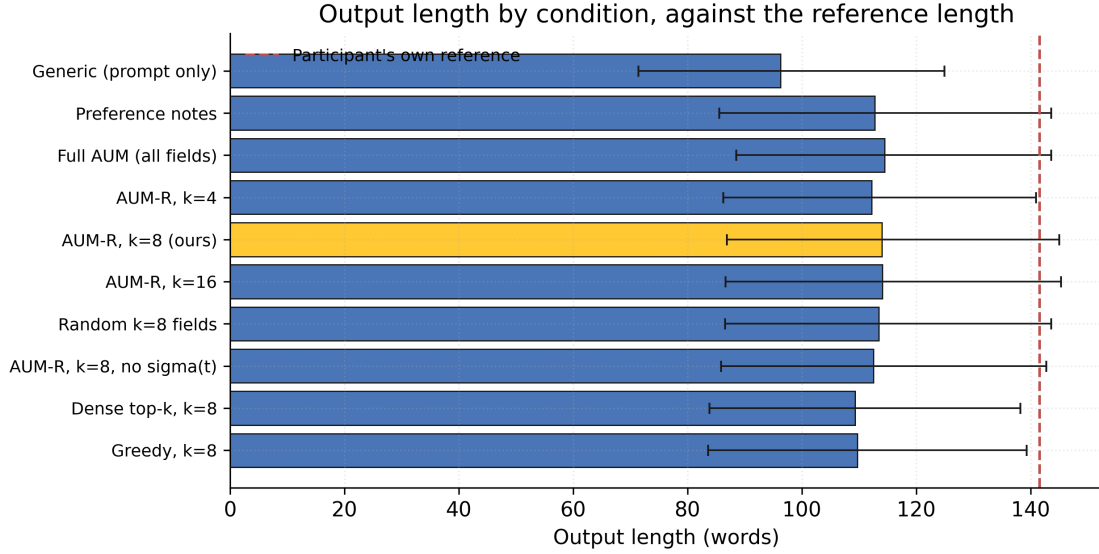


Figure 14: Output length by condition against the participants’ own reference length. Every user-model condition moves output length towards the reference; the generic condition is furthest from it.

9. Discussion

9.1. A context result, and why it is a privacy result

The finding we would put first is not the effect size but the ratio. Eight retrieved fields, 211 tokens, matched the fidelity of injecting the entire 32-field user model, 915 tokens. Nothing in the study suggests that the remaining twenty-four fields were doing work that the eight did not.

For efficiency this is a modest saving. For architecture it is the whole argument. Section 4 committed AUM to local residence: the store sits on the user’s device or in a user-controlled vault, never in a third-party log. That commitment is easy to state and hard to keep if the whole model must be transmitted on every complex request, because a model that is transmitted in full on every turn has, for practical purposes, been uploaded. A budgeted payload changes the character of the commitment. If eight fields suffice, disclosure per query is bounded by construction, the bound is measurable, and the fields that never leave are disproportionately the inner-shell ones, which are exactly the ones whose exposure carries the greatest cost.

In the vocabulary of contextual integrity (Nissenbaum, 2011), the design question is not whether the user model is secret but whether each flow out of it matches the norms of the context that produced it. A retrieval stage that returns a small task-conditioned slice is a mechanism for making flows contextual, and the size of the slice is the parameter that governs it. This is what we mean by privacy by architecture rather than privacy by policy: the former is a property one can measure in tokens, and we measured it.

There is a second consequence for scrutability. Eight short natural-language strings can be shown to a user before they are sent; thirty-two cannot, and a JSON blob certainly cannot. The scrutable-personalisation literature has long argued that users should be able to inspect and correct the model a system holds of them (Kay and Kummerfeld, 2012; Czarkowski and Kay, 2002; Bakalov et al., 2010), and evidence that transparency and control change acceptance and behaviour is not new (Cramer et al., 2008; Harper et al., 2015). The obstacle has usually been that the model is too

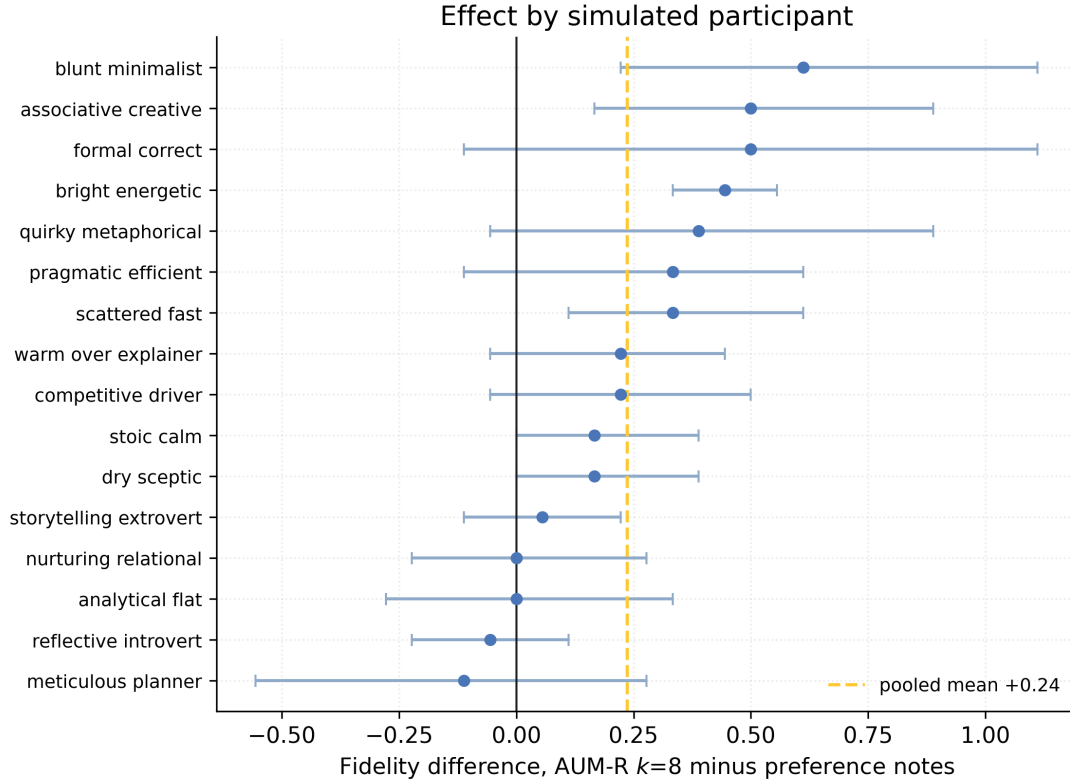


Figure 15: Effect of AUM-R at $k = 8$ over preference notes, by simulated participant, with 95% bootstrap intervals over that participant’s six tasks. The dashed line is the pooled mean.

large or too statistical to show. A budgeted payload of readable sentences removes that obstacle for the per-request case: a user can be shown exactly what the assistant was told about them for this draft.

9.2. What the null results bound, and what they do not

Four controls returned null, and Section 8 explained why. We think the joint reading is more interesting than any of them individually.

The positive results establish that injecting a structured, task-plausible slice of a user model changes generated style substantially relative to both a generic baseline and the flat preference notes deployed systems store. The nulls establish that, at this scale, nothing about *which* slice or about *how hard one searches for it* adds anything further. Together they locate the effect in the representation rather than in the retrieval machinery. That is a narrower claim than the one the architecture invites, and it has a practical implication we would act on: a first deployment should use forward greedy over a well-specified schema, and should not spend engineering effort on the selection algorithm until the schema is larger and the redundancy penalty is doing real work.

Two boundaries on the nulls deserve emphasis. First, they are nulls at an operating point with a pool of about seventeen fields, and Section 8 shows the optimisation landscape is genuinely flat there: greedy is within 0.1% of the exhaustive optimum, so there was nothing for a better search to recover. They are not evidence that selection is irrelevant at the full sixty-field specification, and we

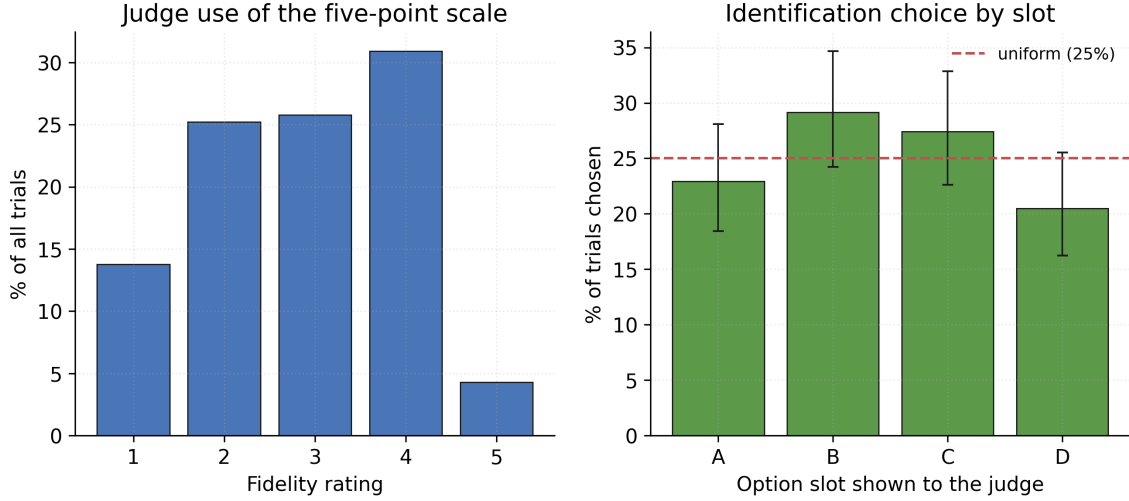


Figure 16: Judge diagnostics. Left: use of the five-point scale across all trials. Right: identification choices by option slot, against the uniform 25% expectation, with Wilson intervals.

do not present them as such. Second, a null against random selection within a $\sigma(t)$ -filtered pool is not a null against random selection from the whole model. The pool was already restricted to shells the task touches, so “random” here means random among plausible fields, which is a considerably stronger baseline than it sounds.

9.3. Personalisation is worth most to the people the default serves least

The result with the clearest design consequence is the one in Section 7.8: the benefit of the user model was largest exactly for those participants whose voice the un-personalised assistant reproduced worst ($\rho = -0.61$, $p = 0.013$ at participant level). The blunt minimalist gained +0.611; the meticulous planner, whose numbered and dated register a default assistant already produces, gained nothing.

Read as a statistic this is a moderate correlation on sixteen points that did not replicate on the automatic style metric. Read as a design observation it says something sharper. A default assistant voice is not neutral. It is a particular register, and being well served by an un-personalised assistant means happening to write the way that register writes. Personalisation, on this reading, is not a uniform improvement distributed evenly across users; it is a correction whose size depends on how far a person sits from the model’s prior.

This connects to a finding that motivates the whole project. [Jakesch et al. \(2023\)](#) showed that co-writing with an opinionated language model shifts users’ own expressed views. If a default assistant voice can move what a person says, then users whose register is furthest from the default face the largest pull towards it, and are also the users for whom a model of their own voice does the most work. Personalisation in that framing is not a convenience feature but a countermeasure, and the task-level pattern in Table 12 says the same thing at a different granularity: the effect was near zero on the deadline-extension email, the most conventionalised genre in the battery, where the assistant’s default already is the correct answer.

9.4. Implications for the design of personalised assistants

Four implications follow that we would apply to a system.

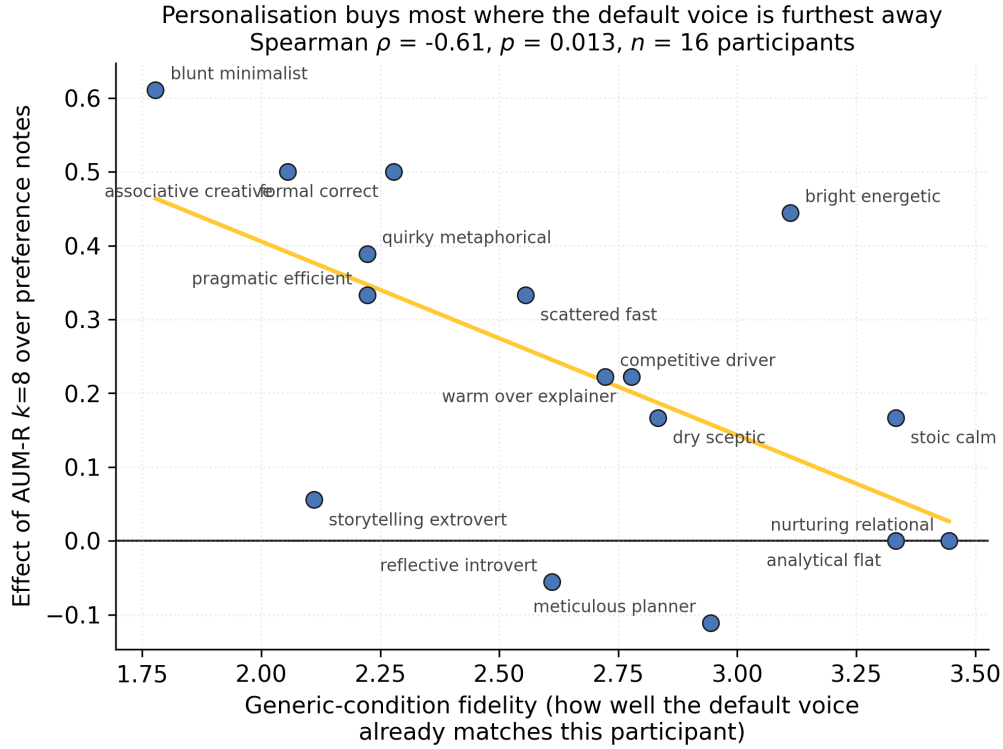


Figure 17: The treatment effect against how well the un-personalised assistant already matches each participant. Points are the sixteen simulated participants; the line is an ordinary least squares fit.

1. **Store the person, retrieve the preference.** The unit of storage should be the stable layer, and the situational preference should be derived at generation time. This is what makes an unseen task addressable and a stale preference re-interpretable rather than merely old.
2. **Budget the payload, and show it.** A small task-conditioned payload costs nothing measurable in quality, bounds disclosure per request, and is small enough to display. A system that shows the user the eight sentences it used has made scrutability concrete rather than aspirational.
3. **Route simple requests around the model.** The classifier stage routed factual lookups to bypass with 98.8% accuracy. This is cheap to implement and eliminates a class of disclosure that buys nothing, and we would regard omitting it as a design error rather than a missed optimisation.
4. **Expect uneven benefit, and design for the tail.** If the gain concentrates on users whose voice the default serves worst, then average-case evaluation will systematically understate the value of personalisation to the users who need it most. Reporting a mean effect across a user population is not sufficient; the distribution is the result.

9.5. Limitations

Simulated participants.. The most important limitation is the one stated first in Section 6.1. Every participant here is language-model-simulated. A judge model rating on behalf of a persona description has an easier task than a person recognising their own voice, and the absolute fidelity

Table 15: Effect of AUM-R at $k = 8$ over preference notes, by simulated participant, ordered by effect size. “Generic” is that participant’s mean fidelity in the prompt-only condition, which measures how well the default voice already matches them.

Participant	Generic	Δ	95% CI	Favouring
blunt minimalist	1.78	+0.611	[+0.222, +1.111]	83.3%
associative creative	2.06	+0.500	[+0.167, +0.889]	66.7%
formal correct	2.28	+0.500	[−0.111, +1.111]	66.7%
bright energetic	3.11	+0.444	[+0.333, +0.556]	100.0%
quirky metaphorical	2.22	+0.389	[−0.056, +0.889]	50.0%
pragmatic efficient	2.22	+0.333	[−0.111, +0.611]	83.3%
scattered fast	2.56	+0.333	[+0.111, +0.611]	66.7%
warm over-explainer	2.78	+0.222	[−0.056, +0.444]	66.7%
competitive driver	2.72	+0.222	[−0.056, +0.500]	50.0%
stoic calm	3.33	+0.167	[0.000, +0.389]	33.3%
dry sceptic	2.83	+0.167	[0.000, +0.389]	33.3%
storytelling extrovert	2.11	+0.056	[−0.111, +0.222]	33.3%
nurturing relational	3.44	0.000	[−0.222, +0.278]	16.7%
analytical flat	3.33	0.000	[−0.278, +0.333]	33.3%
reflective introvert	2.61	−0.056	[−0.222, +0.111]	16.7%
meticulous planner	2.94	−0.111	[−0.556, +0.278]	16.7%

Table 16: Relative gap to the best solution found, in percent, averaged over k and 25 instances per cell, at $\lambda = 0.15$. Lower is better. Forward greedy is within 0.1% of the best solution at every instrument size; the swarm at the study’s fixed budget degrades steeply.

n	Greedy	Dense top- k	AFSA study	AFSA scaled	AFSA warm
16	0.10	0.65	0.44	0.04	0.00
32	0.08	0.46	5.66	0.34	0.00
64	0.01	0.35	20.94	0.57	0.01
128	0.07	0.47	33.51	5.46	0.05
256	0.03	0.53	43.23	12.59	0.01

means sit near the middle of a five-point scale for every condition, including the best. The critical literature on synthetic participants (Bisbee et al., 2024; Wang et al., 2025) applies to this study in full, and no result here should be transferred to a claim about human users.

The personas are not a sample.. Sixteen personas written to be distinguishable are a friendlier population than any real one. The idiolect markers were authored, so the style differences the study measures are differences an author put there. In particular, conclusions about *which* users benefit should not be transferred to demographic groups, since the personas carry an author’s assumptions about how people with a given trait profile write.

Judge reliability is moderate.. ICC(2,1) of 0.57 across seeds is moderate rather than good (Koo and Li, 2016). Agreement within one scale point was 86%, so the judge is consistent about direction and imprecise about magnitude, which is adequate for paired contrasts and inadequate for absolute claims. Position bias was not detected but the test was close to threshold ($p = 0.061$).

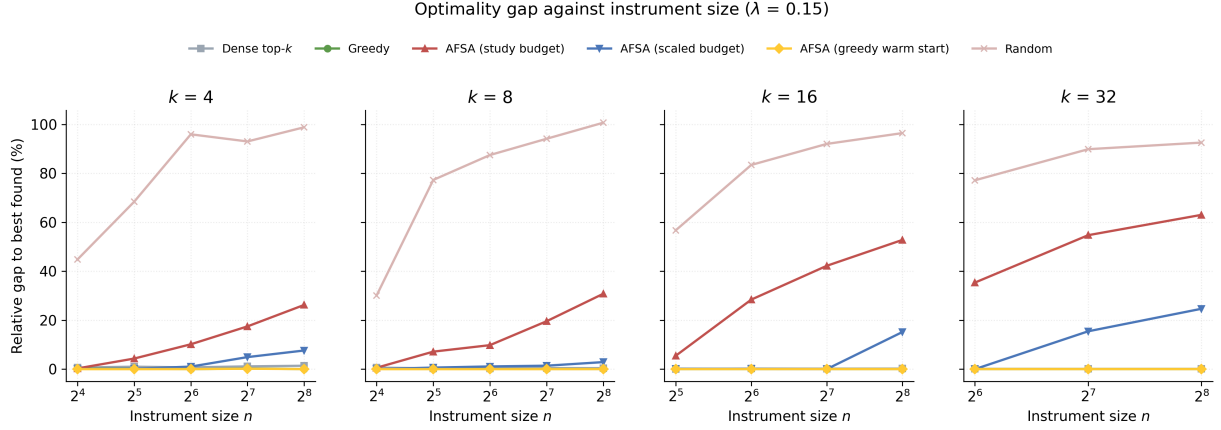


Figure 18: Relative optimality gap against instrument size, one panel per budget k , at $\lambda = 0.15$. Greedy and the warm-started swarm are indistinguishable from the best solution found at every scale; the fixed-budget swarm used in the simulation study degrades steeply.

Table 17: Objective evaluations at $k = 8$, $\lambda = 0.15$, averaged over 25 instances. Dense top- k and random require a single evaluation. Exhaustive search is tractable only at $n = 16$.

n	Greedy	AFSA study	AFSA scaled	AFSA warm	Exhaustive
16	101	417	981	1,083	12,871
32	229	468	3,112	3,375	n/a
64	485	521	6,556	7,112	n/a
128	997	479	10,933	11,926	n/a
256	2,021	375	13,963	16,002	n/a

Acquisition is unspecified.. AUM specifies the fields a user representation should contain and not how they are populated. Self report is unreliable for inner-shell entries (Vazire, 2010; Connelly and Ones, 2010; Baumeister et al., 2007); passive observation raises the privacy concerns the architecture exists to avoid; and clinical assessment does not scale. We treat this as an open problem on which AUM as a target representation is agnostic, but it is the largest gap between this paper and a deployable system.

Update is unspecified.. Fields carry a nominal update period τ , but no mechanism is given for detecting that a field has become stale or for revising it without re-running the whole instrument. Given that the motivating example in Section 1 is precisely a staleness failure, this is a conspicuous omission.

Text only, and one language.. The specification is text-based. Many personality cues are non-verbal: tone of voice, facial expression, pacing, response latency. The study also runs entirely in English, and the personality frameworks AUM draws on were developed predominantly in specific populations, so cross-cultural validity of the inner tiers is untested (McCrae, 2002).

Single generator and judge.. One generator model and one judge model were used. Whether the effect sizes transfer across model families is unknown, and a judge from a single family may carry systematic preferences that a panel would average out.

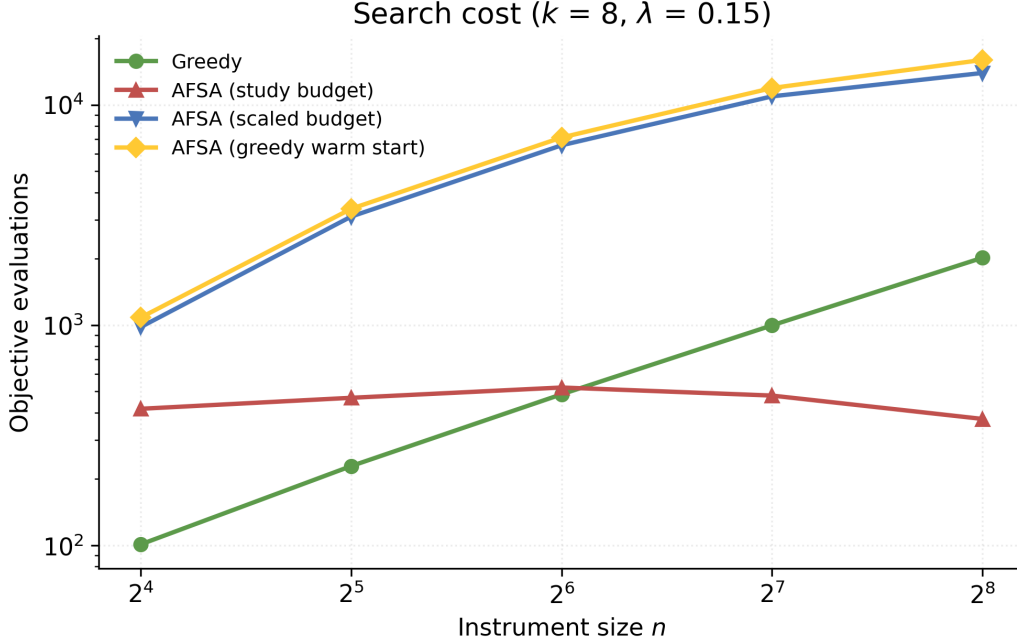


Figure 19: Search cost in objective evaluations at $k = 8$, $\lambda = 0.15$. The fixed-budget swarm spends a roughly constant amount regardless of instrument size, which is exactly why its solution quality collapses as n grows.

Table 18: Relative gap to the best solution found, in percent, against the redundancy weight λ at $n = 64$, $k = 8$. At $\lambda = 0$ the objective is separable and sorting is optimal.

λ	Dense top- k	Greedy	AFSA scaled	AFSA warm
0.00	0.00	0.00	1.00	0.00
0.15	0.32	0.00	1.11	0.00
0.30	1.51	0.35	0.87	0.04
0.60	7.68	1.73	1.76	0.03

Lifespan and capacity. AUM as specified assumes an adult user with an articulable identity. It does not represent young children, or persons with significant cognitive impairment, and we do not think it should be extended to them without a separate ethical analysis.

9.6. Future work

The necessary next study is the human one this simulation substitutes for. Its shape is determined by what the simulation cannot answer: participants populate their own AUM, write their own references for the same six tasks, and rate outputs from the same conditions blind, with the primary measure being self-rated fidelity and forced-choice identification of their own style. Section 10 states the conditions under which we think that study can be run.

Three further directions follow from the results rather than from the gaps. First, acquisition: whether an AUM can be populated incrementally from ordinary interaction without the passive observation the architecture is meant to avoid, perhaps by asking the user a small number of well-chosen questions at moments when the answer is cheap to give. Second, update: detecting staleness

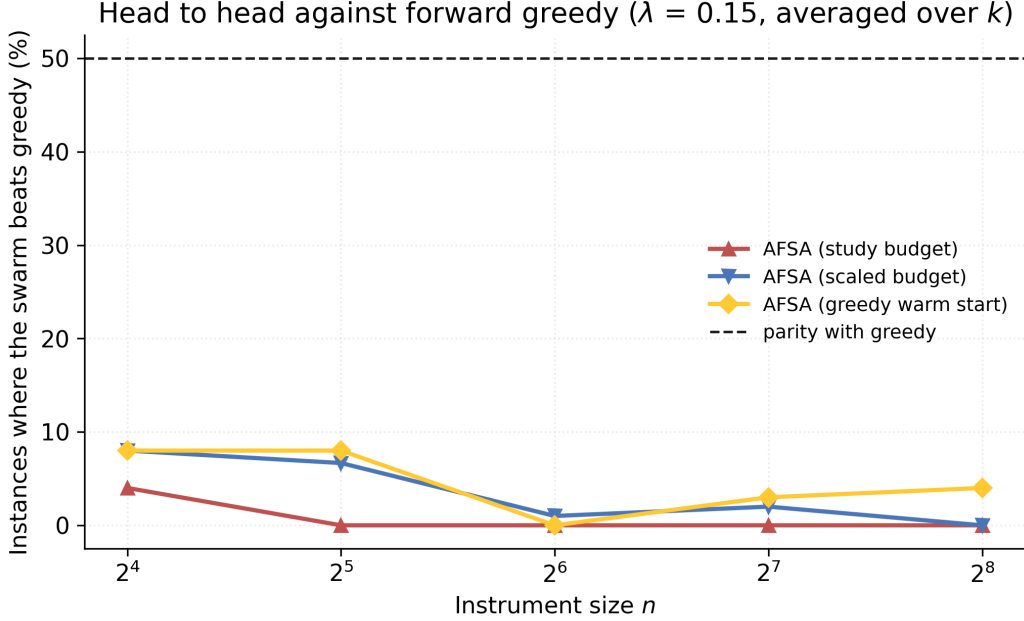


Figure 20: Head-to-head strict win rate against forward greedy at $\lambda = 0.15$, averaged over k . Strict wins are rare because ties are common: greedy and the warm-started swarm frequently return the same subset.

by watching for conflict between a stored field and new evidence, which is what the cross-shell internal-conflicts entry exists to support. Third, scale: whether the selection nulls of Section 7.4 persist at the full sixty-field specification, where Section 8 suggests the optimisation landscape is no longer flat.

10. Ethical Considerations

A structured user model that records inner-shell content such as fears, difficult episodes and core values is among the most sensitive artefacts a person could hold. Building one creates risks that a preference list does not, and we do not think those risks are adequately answered by noting that the model is opt-in.

Three binding constraints.. We bind AUM to three. **Local residence:** the store sits on the user’s device or in a user-controlled vault, never in a third-party log, and only the retrieved payload leaves it. Section 7 makes this quantitative rather than aspirational: a payload of eight fields reaches the fidelity of the whole model, so budgeted retrieval bounds what any single query can disclose at no measured cost in quality. **User authorship:** every field is inspectable, editable and deletable, and eight short strings are few enough to display before they are sent. **Purpose restriction:** AUM is intended for task-style personalisation under the user’s direction, not for profiling, advertising, surveillance, or persuasion against the user’s interest.

The misuse the architecture cannot prevent.. A high-resolution model of a person is valuable to anyone who wants to influence that person, and an operator who ignores the local-residence constraint gains a profiling asset of a kind that does not currently exist at scale. Our own results

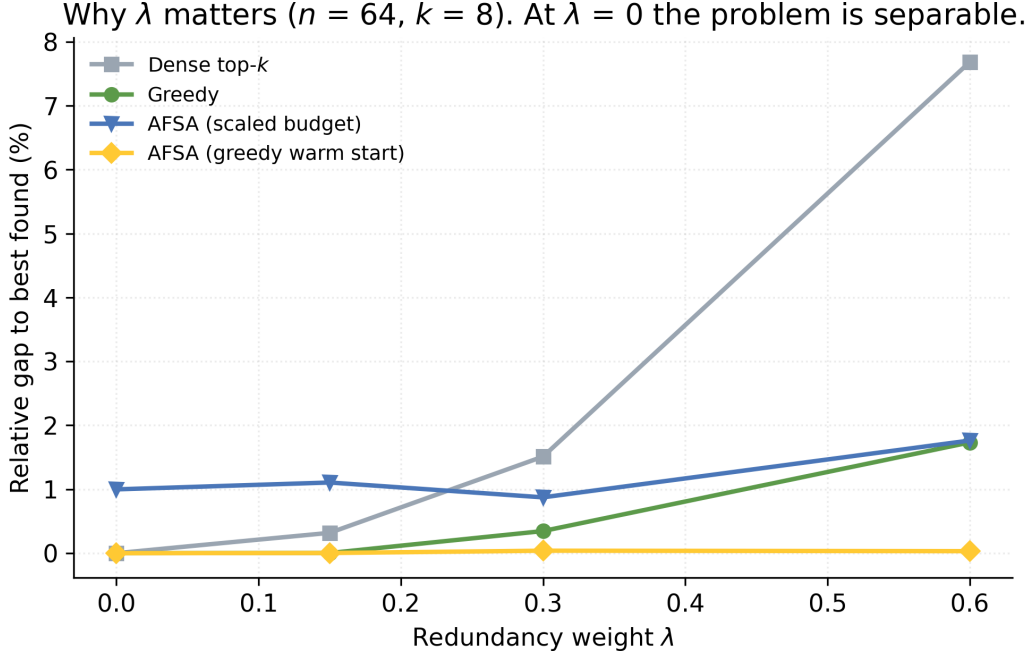


Figure 21: Effect of the redundancy weight at $n = 64, k = 8$. At $\lambda = 0$ the problem is a sort; as λ rises, relevance-only selection loses ground and combinatorial search starts to earn its cost.

sharpen this rather than softening it: Section 7.8 found the effect largest for participants whose voice the default serves worst, and those are also the people for whom an accurate model would be most identifying. Users of LLM-based conversational agents already disclose a great deal and reason about the risks only partially (Zhang et al., 2024b), and language models can leak personal information present in training data (Huang et al., 2022).

We do not think this argues against specifying the representation. Opaque, implicit user models already sit inside large-scale systems with weaker user control, and an explicit schema at least makes the contents auditable and contestable. But mitigation has to be architectural, in where the data lives, as well as procedural, in who may query it and for what, and a paper that proposes the representation without saying so would be incomplete.

Human participants.. No human participants were involved in this work and no human-subjects data was collected. Every participant in Sections 6 and 7 is language-model-simulated and every persona is fictional and declared in code (Appendix Appendix C). The human study described in Section 9.6 would require informed consent, the right to withdraw with deletion of all collected AUM fields, and institutional ethics approval, since participants would disclose identity-level content about themselves. Inner-shell fields should be optional by default and skippable without penalty, and participants should be able to review their populated instrument before any of it is used.

A concern specific to simulation.. Personas written by an author carry that author’s assumptions about how people of a given trait profile write. Ours are deliberately varied but are not a sample of any population. Conclusions about which users benefit from personalisation should not be transferred to demographic groups on the basis of this study, and we have tried to phrase Section 7.8 so that it cannot be read that way.

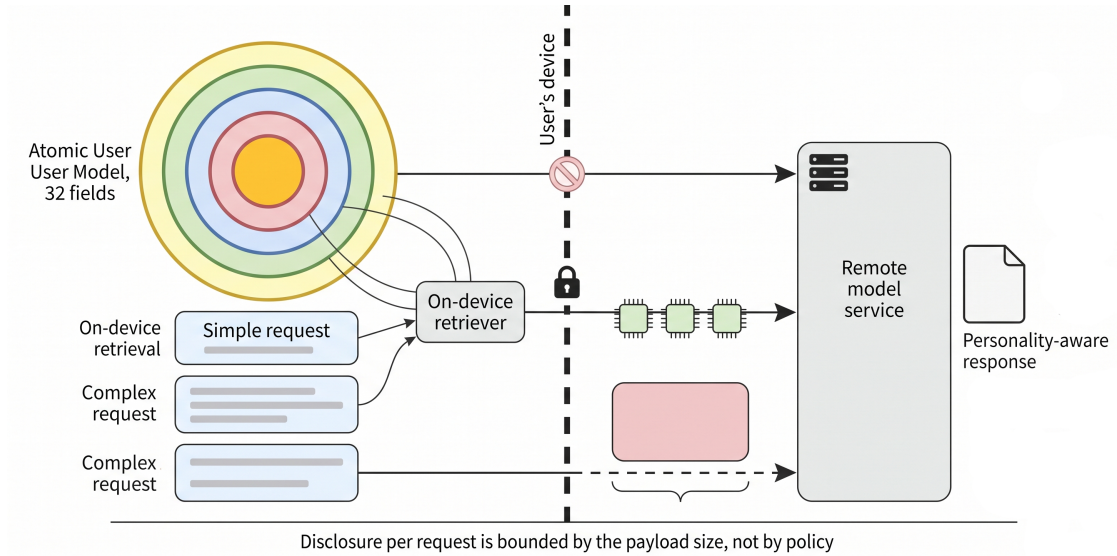


Figure 22: What crosses the device boundary. The full 32-field model, 915 tokens in the study, never leaves the device. A simple request leaves nothing. A complex request sends a payload of eight fields, 211 tokens, which the user can be shown before it is sent. The bound on disclosure per query is a measured quantity rather than a policy promise.

11. Conclusion

Personalisation of language-model assistants currently stores what a user has done. We argued that it should store the person from whom those doings come, and that the two are not interchangeable because personality is comparatively stable while preferences are situational. We named the phenomenon that makes the difference visible, personality seepage, in which a prompt carries a fingerprint the assistant mirrors without being able to read. We specified the Atomic User Model, a five-tier structured and human-readable representation whose metaphor encodes differential stability and differential observability, and a personality-aware retrieval pipeline that treats it as an index rather than a prompt prefix.

In a simulation study with sixteen language-model-simulated participants, six style-sensitive tasks and three seeds, retrieving eight fields matched the style fidelity of injecting the whole user model at 23% of the context, improved on the flat preference notes deployed systems store by 0.24 points on a five-point scale, and raised forced-choice identification of the participant’s own voice from 14.9% to 42.7% against 25% chance. Four controls returned null, and a synthetic scaling study of the retriever explained why: at the operating point used, forward greedy is already within 0.1% of the exhaustive optimum, so there was nothing for a population method to recover. The effect lives in the representation, not in the search over it.

The result we would carry forward is neither the effect size nor the null. It is that a small, readable, task-conditioned payload is enough, because that is what makes a user model something a person can keep on their own device and read before it is sent. And it is that the benefit was largest for the participants whose voice the default assistant served worst, which suggests that personalisation is least valuable to the users who resemble the model’s prior and most valuable to everyone else.

CRediT authorship contribution statement

B. Sankar: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing, original draft, Writing, review and editing, Supervision, Project administration. **Deepthika S:** Methodology, Software, Validation, Investigation, Data curation, Writing, review and editing. **Pawni Yadav:** Methodology, Software, Validation, Investigation, Visualization, Writing, review and editing. **Amogh A S:** Software, Validation, Data curation, Visualization, Writing, review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI in the writing process

The studies reported in Sections 6 to 8 use large language models as the object of study and as simulated participants; this use is described in full in those sections. In preparing this manuscript the authors additionally used a large language model to assist with drafting and editing of prose. The authors reviewed and edited all content and take full responsibility for the content of the publication.

Data and code availability

The simulation code, the persona declarations, every prompt, and the run manifest and ledger, all per-trial records, the analysis scripts, and the scaling benchmark are available at <https://github.com/REPOSITORY-URL-HERE>. Each number reported in Sections 7 and 8 is reproducible from the released artefacts and a seed.

Funding declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Agrawal, R., Gollapudi, S., Halverson, A., Jeong, S., 2009. Diversifying search results, in: Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM. pp. 5–14. doi:[10.1145/1498759.1498766](https://doi.org/10.1145/1498759.1498766).
- Alhafni, B., Kulkarni, V., Kumar, D., Raheja, V., 2024. Personalized text generation with fine-grained linguistic control, in: Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), Association for Computational Linguistics. pp. 88–101. doi:[10.18653/v1/2024.personalize-1.8](https://doi.org/10.18653/v1/2024.personalize-1.8).
- Altman, I., Taylor, D.A., 1973. Social Penetration: The Development of Interpersonal Relationships. Holt, Rinehart and Winston, New York.
- Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review 111, 1036–1060. doi:[10.1037/0033-295X.111.4.1036](https://doi.org/10.1037/0033-295X.111.4.1036).
- Andrews, N., Bishop, M., 2019. Learning invariant representations of social media users, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics. pp. 1684–1695. doi:[10.18653/v1/D19-1178](https://doi.org/10.18653/v1/D19-1178).

- Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D., 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 337–351. doi:[10.1017/pan.2023.2](https://doi.org/10.1017/pan.2023.2).
- Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 555–596. doi:[10.1162/coli.07-034-R2](https://doi.org/10.1162/coli.07-034-R2).
- Azucar, D., Marengo, D., Settanni, M., 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences* 124, 150–159. doi:[10.1016/j.paid.2017.12.018](https://doi.org/10.1016/j.paid.2017.12.018).
- Bakalov, F., König-Ries, B., Nauerz, A., Welsch, M., 2010. IntrospectiveViews: An interface for scrutinizing semantic user models, in: *User Modeling, Adaptation, and Personalization*. Springer. Lecture Notes in Computer Science, pp. 219–230. doi:[10.1007/978-3-642-13470-8_21](https://doi.org/10.1007/978-3-642-13470-8_21).
- Baumeister, R.F., Vohs, K.D., Funder, D.C., 2007. Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science* 2, 396–403. doi:[10.1111/j.1745-6916.2007.00051.x](https://doi.org/10.1111/j.1745-6916.2007.00051.x).
- Bennett, P.N., White, R.W., Chu, W., Dumais, S.T., Bailey, P., Borisyuk, F., Cui, X., 2012. Modeling the impact of short- and long-term behavior on search personalization, in: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, USA. pp. 185–194. doi:[10.1145/2348283.2348312](https://doi.org/10.1145/2348283.2348312).
- Bisbee, J., Clinton, J.D., Dorff, C., Kenkel, B., Larson, J.M., 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis* 32, 401–416. doi:[10.1017/pan.2024.5](https://doi.org/10.1017/pan.2024.5).
- Bowlby, J., 1969. Attachment. volume 1 of *Attachment and Loss*. Basic Books, New York.
- Boyd, R.L., Pennebaker, J.W., 2017. Language-based personality: a new approach to personality in a digital world. *Current Opinion in Behavioral Sciences* 18, 63–68. doi:[10.1016/j.cobeha.2017.07.017](https://doi.org/10.1016/j.cobeha.2017.07.017).
- Brusilovsky, P., Millán, E., 2007. User models for adaptive hypermedia and adaptive educational systems, in: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer. volume 4321 of *Lecture Notes in Computer Science*, pp. 3–53. doi:[10.1007/978-3-540-72079-9_1](https://doi.org/10.1007/978-3-540-72079-9_1).
- Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries, in: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM. pp. 335–336. doi:[10.1145/290941.291025](https://doi.org/10.1145/290941.291025).
- Carpenter, A., Greene, K., 2015. Social penetration theory, in: Berger, C.R., Roloff, M.E. (Eds.), *The International Encyclopedia of Interpersonal Communication*. Wiley, pp. 1–4. doi:[10.1002/9781118540190.wbeic160](https://doi.org/10.1002/9781118540190.wbeic160).
- Chen, M., Suresh, A.T., Mathews, R., Wong, A., Allauzen, C., Beaufays, F., Riley, M., 2019. Federated learning of n-gram language models, in: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics. pp. 121–130. doi:[10.18653/v1/K19-1012](https://doi.org/10.18653/v1/K19-1012).
- Chhetri, G., Das, S., Chowdhury, T.I., 2026. SPARK: Search personalization via agent-driven retrieval and knowledge-sharing, in: *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, New York, NY, USA. pp. 84–92. doi:[10.1145/3779211.3793173](https://doi.org/10.1145/3779211.3793173).
- Chiang, C.H., Lee, H.y., 2023. Can large language models be an alternative to human evaluations?, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics. pp. 15607–15631. doi:[10.18653/v1/2023.acl-long.870](https://doi.org/10.18653/v1/2023.acl-long.870).
- Cho, I., Wang, D., Takahashi, R., Saito, H., 2022. A personalized dialogue generator with implicit user persona detection, in: *Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics*, Gyeongju, Republic of Korea. pp. 367–377. URL: <https://aclanthology.org/2022.coling-1.29/>.
- Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I., 2008. Novelty and diversity in information retrieval evaluation, in: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM. pp. 659–666. doi:[10.1145/1390334.1390446](https://doi.org/10.1145/1390334.1390446).
- Connelly, B.S., Ones, D.S., 2010. An other perspective on personality: Meta-analytic integration of observers’ accuracy and predictive validity. *Psychological Bulletin* 136, 1092–1122. doi:[10.1037/a0021212](https://doi.org/10.1037/a0021212).
- Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B., 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 455–496. doi:[10.1007/s11257-008-9051-3](https://doi.org/10.1007/s11257-008-9051-3).
- Czarkowski, M., Kay, J., 2002. A scrutable adaptive hypertext, in: *Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer. Lecture Notes in Computer Science, pp. 384–387. doi:[10.1007/3-540-47952-X_43](https://doi.org/10.1007/3-540-47952-X_43).
- Dou, Z., Song, R., Wen, J.R., 2007. A large-scale evaluation and analysis of personalized search strategies, in: *Proceedings of the 16th International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, USA. pp. 581–590. doi:[10.1145/1242572.1242651](https://doi.org/10.1145/1242572.1242651).

- Epstein, D.A., Ping, A., Fogarty, J., Munson, S.A., 2015. A lived informatics model of personal informatics, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM. pp. 731–742. doi:[10.1145/2750858.2804250](https://doi.org/10.1145/2750858.2804250).
- Erikson, E.H., 1968. Identity: Youth and Crisis. W. W. Norton, New York.
- Gero, K.I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D.R., Campbell, M., Kumaravel, S., Zhang, W., 2020. Mental models of AI agents in a cooperative game setting, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, ACM. pp. 1–12. doi:[10.1145/3313831.3376316](https://doi.org/10.1145/3313831.3376316).
- Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, New York.
- Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G., 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality* 40, 84–96. doi:[10.1016/j.jrp.2005.08.007](https://doi.org/10.1016/j.jrp.2005.08.007).
- Gosling, S.D., Rentfrow, P.J., Swann, W.B., 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 504–528. doi:[10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1).
- Goyal, M., Tawde, P., 2022. A research attempt to predict and model personalities through users’ social media details, in: 2022 IEEE Bombay Section Signature Conference (IBSSC), IEEE. pp. 1–6. doi:[10.1109/IBSSC56953.2022.10037272](https://doi.org/10.1109/IBSSC56953.2022.10037272).
- Harper, F.M., Xu, F., Kaur, H., Condiff, K., Chang, S., Terveen, L., 2015. Putting users in control of their recommendations, in: Proceedings of the 9th ACM Conference on Recommender Systems, ACM. pp. 3–10. doi:[10.1145/2792838.2800179](https://doi.org/10.1145/2792838.2800179).
- Huang, F., Bei, Y., Yang, Z., Jiang, J., Chen, H., Shen, Q., Wang, S., Karray, F., Yu, P.S., 2025. Large language model simulator for cold-start recommendation, in: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 261–270. doi:[10.1145/3701551.3703546](https://doi.org/10.1145/3701551.3703546).
- Huang, J., Shao, H., Chang, K.C.C., 2022. Are large pre-trained language models leaking your personal information?, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics. pp. 2038–2047. doi:[10.18653/v1/2022.findings-emnlp.148](https://doi.org/10.18653/v1/2022.findings-emnlp.148).
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M., 2023. Co-writing with opinionated language models affects users’ views, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, ACM. pp. 1–15. doi:[10.1145/3544548.3581196](https://doi.org/10.1145/3544548.3581196).
- Jeromela, J., 2022. Scrutability of intelligent personal assistants, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, ACM. pp. 335–340. doi:[10.1145/3503252.3534355](https://doi.org/10.1145/3503252.3534355).
- Ji, Y., Wu, W., Zheng, H., Hu, Y., Chen, X., He, L., 2023. Is ChatGPT a good personality recognizer? A preliminary study. URL: <https://arxiv.org/abs/2307.03952>, doi:[10.48550/arXiv.2307.03952](https://doi.org/10.48550/arXiv.2307.03952), arXiv:2307.03952.
- Jiang, H., Wu, Q., Lin, C.Y., Yang, Y., Qiu, L., 2023. LLMingua: Compressing prompts for accelerated inference of large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 13358–13376. doi:[10.18653/v1/2023.emnlp-main.825](https://doi.org/10.18653/v1/2023.emnlp-main.825).
- Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.Y., Yang, Y., Qiu, L., 2024. LongLLMingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics. pp. 1658–1677. doi:[10.18653/v1/2024.acl-long.91](https://doi.org/10.18653/v1/2024.acl-long.91).
- Kang, W.C., Ni, J., Mehta, N., Sathiamoorthy, M., Hong, L., Chi, E., Cheng, D.Z., 2023. Do LLMs understand user preferences? evaluating LLMs on user rating prediction. URL: <https://arxiv.org/abs/2305.06474>, doi:[10.48550/arXiv.2305.06474](https://doi.org/10.48550/arXiv.2305.06474), arXiv:2305.06474.
- Kay, J., Kummerfeld, B., 2012. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems* 2, 1–42. doi:[10.1145/2395123.2395129](https://doi.org/10.1145/2395123.2395129).
- Kobsa, A., 2001. Generic user modeling systems. *User Modeling and User-Adapted Interaction* 11, 49–63. doi:[10.1023/A:1011187500863](https://doi.org/10.1023/A:1011187500863).
- Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15, 155–163. doi:[10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012).
- Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 5802–5805. doi:[10.1073/pnas.1218772110](https://doi.org/10.1073/pnas.1218772110).
- Krippendorff, K., 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research* 30, 411–433. doi:[10.1111/j.1468-2958.2004.tb00738.x](https://doi.org/10.1111/j.1468-2958.2004.tb00738.x).
- Kumar, I., Viswanathan, S., Yerra, S., Salemi, A., Rossi, R.A., Dernoncourt, F., Deilamsalehy, H., Chen, X., Zhang,

- R., Agarwal, S., Lipka, N., Nguyen, C.V., Nguyen, T.H., Zamani, H., 2024. LongLaMP: A benchmark for personalized long-form text generation. URL: <https://arxiv.org/abs/2407.11016>, doi:10.48550/arXiv.2407.11016, arXiv:2407.11016.
- Laird, J.E., 2012. The Soar Cognitive Architecture. MIT Press, Cambridge, MA.
- Lee, M., Gero, K.I., Chung, J.J.Y., Buckingham Shum, S., Raheja, V., Shen, H., Venugopalan, S., Wambsganss, T., Zhou, D., Alghamdi, E.A., August, T., Bhat, A., Choksi, M.Z., Dutta, S., Guo, J.L.C., Hoque, M.N., Kim, Y., Knight, S., Neshaei, S.P., Shibani, A., Shrivastava, D., Shroff, L., Sergeyuk, A., Stark, J., Sterman, S., Wang, S., Bosselut, A., Buschek, D., Chang, J.C., Chen, S., Kreminski, M., Park, J., Pea, R., Rho, E.H.R., Shen, Z., Siangliulue, P., 2024. A design space for intelligent and interactive writing assistants, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM. pp. 1–35. doi:10.1145/3613904.3642697.
- Lee, M., Liang, P., Yang, Q., 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, ACM. pp. 1–19. doi:10.1145/3491102.3502030.
- Li, I., Dey, A., Forlizzi, J., 2010. A stage-based model of personal informatics systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. pp. 557–566. doi:10.1145/1753326.1753409.
- Li, X.L., Shao, Z.J., Qian, J.X., 2002. An optimizing method based on autonomous animats: Fish-swarm algorithm. Systems Engineering - Theory & Practice 22, 32–38. No DOI is registered for this record; verified on the publisher's article page.
- Li, Y., Dong, B., Guerin, F., Lin, C., 2023. Compressing context to enhance inference efficiency of large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 6342–6353. doi:10.18653/v1/2023.emnlp-main.391.
- Liao, Y., Deng, Y., Jiang, T., Ren, J., 2026. Personalizing large language models with user profile memory, in: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, Association for Computing Machinery, New York, NY, USA. pp. 3051–3061. doi:10.1145/3770855.3817930.
- Lin, Y., Chen, L., Ali, A., Nugent, C., Cleland, I., Li, R., Ding, J., Ning, H., 2024. Human digital twin: a survey. Journal of Cloud Computing 13, 131. doi:10.1186/s13677-024-00691-z.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C., 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, Association for Computational Linguistics. pp. 2511–2522. doi:10.18653/v1/2023.emnlp-main.153.
- Luger, E., Sellen, A., 2016. “like having a really bad PA”: The gulf between user expectation and experience of conversational agents, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM. pp. 5286–5297. doi:10.1145/2858036.2858288.
- Mahmood, A., Wang, J., Yao, B., Wang, D., Huang, C.M., 2025. User interaction patterns and breakdowns in conversing with LLM-powered voice assistants. International Journal of Human-Computer Studies 195, 103406. doi:10.1016/j.ijhcs.2024.103406.
- Man, H., Nguyen, T.H., 2024. Counterfactual augmentation for robust authorship representation learning, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM. pp. 2347–2351. doi:10.1145/3626772.3657956.
- McCrae, R.R., 2002. Cross-cultural research on the five-factor model of personality. Online Readings in Psychology and Culture 4. doi:10.9707/2307-0919.1038.
- McCrae, R.R., Costa, P.T., 1989. Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. Journal of Personality 57, 17–40. doi:10.1111/j.1467-6494.1989.tb00759.x.
- McCrae, R.R., John, O.P., 1992. An introduction to the five-factor model and its applications. Journal of Personality 60, 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x.
- Mir, R., Felbo, B., Obradovich, N., Rahwan, I., 2019. Evaluating style transfer for text, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics. pp. 495–504. doi:10.18653/v1/N19-1049.
- Nemhauser, G.L., Wolsey, L.A., Fisher, M.L., 1978. An analysis of approximations for maximizing submodular set functions i. Mathematical Programming 14, 265–294. doi:10.1007/BF01588971.
- Neshat, M., Sepidnam, G., Sargolzaei, M., Toosi, A.N., 2014. Artificial fish swarm algorithm: a survey of the state-of-the-art, hybridization, combinatorial and indicative applications. Artificial Intelligence Review 42, 965–997. doi:10.1007/s10462-012-9342-2. published online 2012; print issue 2014.
- Ngo, T., Kunkel, J., Ziegler, J., 2020. Exploring mental models for transparent and controllable recommender systems: A qualitative study, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, ACM. pp. 183–191. doi:10.1145/3340631.3394841.
- Nguyen, B.H., Xue, B., Zhang, M., 2020. A survey on swarm intelligence approaches to feature selection in data

- mining. *Swarm and Evolutionary Computation* 54, 100663. doi:[10.1016/j.swevo.2020.100663](https://doi.org/10.1016/j.swevo.2020.100663).
- Nissenbaum, H., 2011. A contextual approach to privacy online. *Daedalus* 140, 32–48. doi:[10.1162/daed_a_00113](https://doi.org/10.1162/daed_a_00113).
- Ostheimer, P., Nagda, M.K., Kloft, M., Fellenz, S., 2023. A call for standardization and validation of text style transfer evaluation, in: *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics. pp. 10791–10815. doi:[10.18653/v1/2023.findings-acl.687](https://doi.org/10.18653/v1/2023.findings-acl.687).
- Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S.G., Stoica, I., Gonzalez, J.E., 2023. MemGPT: Towards LLMs as operating systems. URL: <https://arxiv.org/abs/2310.08560>, doi:[10.48550/arXiv.2310.08560](https://doi.org/10.48550/arXiv.2310.08560), [arXiv:2310.08560](https://arxiv.org/abs/2310.08560).
- Panickssery, A., Bowman, S., Feng, S., 2024. LLM evaluators recognize and favor their own generations, in: *Advances in Neural Information Processing Systems* 37, pp. 68772–68802. doi:[10.52202/079017-2197](https://doi.org/10.52202/079017-2197).
- Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.P., 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology* 108, 934–952. doi:[10.1037/pspp0000020](https://doi.org/10.1037/pspp0000020).
- Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior, in: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Association for Computing Machinery, New York, NY, USA. pp. 1–22. doi:[10.1145/3586183.3606763](https://doi.org/10.1145/3586183.3606763).
- Park, J.S., Zou, C.Q., Kamphorst, J., Egan, N., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Liang, P., Willer, R., Bernstein, M.S., 2024. LLM agents grounded in self-reports enable general-purpose simulation of individuals. URL: <https://arxiv.org/abs/2411.10109>, doi:[10.48550/arXiv.2411.10109](https://doi.org/10.48550/arXiv.2411.10109), [arXiv:2411.10109](https://arxiv.org/abs/2411.10109).
- Patel, A., Rao, D., Kothary, A., McKeown, K., Callison-Burch, C., 2023. Learning interpretable style embeddings via prompting LLMs, in: *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics. pp. 15270–15290. doi:[10.18653/v1/2023.findings-emnlp.1020](https://doi.org/10.18653/v1/2023.findings-emnlp.1020).
- Pennebaker, J.W., King, L.A., 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77, 1296–1312. doi:[10.1037/0022-3514.77.6.1296](https://doi.org/10.1037/0022-3514.77.6.1296).
- Pittenger, D.J., 2005. Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research* 57, 210–221. doi:[10.1037/1065-9293.57.3.210](https://doi.org/10.1037/1065-9293.57.3.210).
- Pourpanah, F., Wang, R., Lim, C.P., Wang, X.Z., Yazdani, D., 2023. A review of artificial fish swarm algorithms: recent advances and applications. *Artificial Intelligence Review* 56, 1867–1903. doi:[10.1007/s10462-022-10214-4](https://doi.org/10.1007/s10462-022-10214-4). published online 2022; print issue 2023.
- Rao, H., Leung, C., Miao, C., 2023. Can ChatGPT assess human personalities? A general evaluation framework, in: *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics. pp. 1184–1194. URL: <https://aclanthology.org/2023.findings-emnlp.84/>.
- Rapp, A., Cena, F., 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94, 1–17. doi:[10.1016/j.ijhcs.2016.05.006](https://doi.org/10.1016/j.ijhcs.2016.05.006).
- Reza, M., Thomas-Mitchell, J., Dushniku, P., Laundry, N., Williams, J.J., Kuzminykh, A., 2025. Co-writing with AI, on human terms: Aligning research with user demands across the writing process. *Proceedings of the ACM on Human-Computer Interaction* 9, 1–37. doi:[10.1145/3757566](https://doi.org/10.1145/3757566).
- Rivera-Soto, R.A., Miano, O.E., Ordonez, J., Chen, B.Y., Khan, A., Bishop, M., Andrews, N., 2021. Learning universal authorship representations, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 913–919. doi:[10.18653/v1/2021.emnlp-main.70](https://doi.org/10.18653/v1/2021.emnlp-main.70).
- Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin* 126, 3–25. doi:[10.1037/0033-2909.126.1.3](https://doi.org/10.1037/0033-2909.126.1.3).
- Roberts, B.W., Walton, K.E., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin* 132, 1–25. doi:[10.1037/0033-2909.132.1.1](https://doi.org/10.1037/0033-2909.132.1.1).
- Salemi, A., Mysore, S., Bendersky, M., Zamani, H., 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)*, Association for Computational Linguistics. pp. 7370–7392. URL: <https://aclanthology.org/2024.acl-long.399/>.
- Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H., 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* 8, e73791. doi:[10.1371/journal.pone.0073791](https://doi.org/10.1371/journal.pone.0073791).
- Shaikh, O., Sapkota, S., Rizvi, S., Horvitz, E., Park, J.S., Yang, D., Bernstein, M.S., 2025. Creating general user models from computer use, in: *Proceedings of the 38th Annual ACM Symposium on User Interface Software and*

- Technology, Association for Computing Machinery, New York, NY, USA. doi:[10.1145/3746059.3747722](https://doi.org/10.1145/3746059.3747722).
- Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86, 420–428. doi:[10.1037/0033-2909.86.2.420](https://doi.org/10.1037/0033-2909.86.2.420).
- Soto, C.J., John, O.P., 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* 113, 117–143. doi:[10.1037/pspp0000096](https://doi.org/10.1037/pspp0000096).
- Specht, J., Egloff, B., Schmukle, S.C., 2011. Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology* 101, 862–882. doi:[10.1037/a0024950](https://doi.org/10.1037/a0024950).
- Stein, R., Swan, A.B., 2019. Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass* 13, e12434. doi:[10.1111/spc3.12434](https://doi.org/10.1111/spc3.12434).
- Sun, R., 2024. Can a cognitive architecture fundamentally enhance LLMs? or vice versa? URL: <https://arxiv.org/abs/2401.10444>, doi:[10.48550/arXiv.2401.10444](https://doi.org/10.48550/arXiv.2401.10444), arXiv:2401.10444.
- Tan, J., Yang, L., Liu, Z., Liu, Z., Murthy, R., Awalganekar, T.M., Zhang, J., Yao, W., Zhu, M., Kokane, S., Savarese, S., Wang, H., Xiong, C., Heinecke, S., 2025. PersonaBench: Evaluating AI models on understanding personal information through accessing (synthetic) private user data. URL: <https://arxiv.org/abs/2502.20616>, doi:[10.48550/arXiv.2502.20616](https://doi.org/10.48550/arXiv.2502.20616), arXiv:2502.20616.
- Tan, Z., Jiang, M., 2023. User modeling in the era of large language models: Current research and future directions. URL: <https://arxiv.org/abs/2312.11518>, doi:[10.48550/arXiv.2312.11518](https://doi.org/10.48550/arXiv.2312.11518), arXiv:2312.11518.
- Tan, Z., Zeng, Q., Tian, Y., Liu, Z., Yin, B., Jiang, M., 2024. Democratizing large language models via personalized parameter-efficient fine-tuning, in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA. pp. 6476–6491. doi:[10.18653/v1/2024.emnlp-main.372](https://doi.org/10.18653/v1/2024.emnlp-main.372).
- Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 24–54. doi:[10.1177/0261927X09351676](https://doi.org/10.1177/0261927X09351676). published online 2009; print issue 2010.
- Teevan, J., Dumais, S.T., Horvitz, E., 2005. Personalizing search via automated analysis of interests and activities, in: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, USA. pp. 449–456. doi:[10.1145/1076034.1076111](https://doi.org/10.1145/1076034.1076111).
- Vazire, S., 2010. Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology* 98, 281–300. doi:[10.1037/a0017908](https://doi.org/10.1037/a0017908).
- Wang, A., Aggazzotti, C., Kotula, R., Rivera Soto, R., Bishop, M., Andrews, N., 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics* 11, 1416–1431. doi:[10.1162/tacl_a_00610](https://doi.org/10.1162/tacl_a_00610).
- Wang, A., Morgenstern, J., Dickerson, J.P., 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* 7, 400–411. doi:[10.1038/s42256-025-00986-z](https://doi.org/10.1038/s42256-025-00986-z).
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., Sui, Z., 2024a. Large language models are not fair evaluators, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics. pp. 9440–9450. doi:[10.18653/v1/2024.acl-long.511](https://doi.org/10.18653/v1/2024.acl-long.511).
- Wang, T., Tao, M., Fang, R., Wang, H., Wang, S., Jiang, Y.E., Zhou, W., 2024b. AI PERSONA: Towards life-long personalization of LLMs. URL: <https://arxiv.org/abs/2412.13103>, doi:[10.48550/arXiv.2412.13103](https://doi.org/10.48550/arXiv.2412.13103), arXiv:2412.13103.
- Wegmann, A., Schraagen, M., Nguyen, D., 2022. Same author or just same topic? Towards content-independent style representations, in: *Proceedings of the 7th Workshop on Representation Learning for NLP*, Association for Computational Linguistics. pp. 249–268. doi:[10.18653/v1/2022.repl4nlp-1.26](https://doi.org/10.18653/v1/2022.repl4nlp-1.26).
- Xue, B., Zhang, M., Browne, W.N., Yao, X., 2016. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation* 20, 606–626. doi:[10.1109/TEVC.2015.2504420](https://doi.org/10.1109/TEVC.2015.2504420).
- Yarkoni, T., 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44, 363–373. doi:[10.1016/j.jrp.2010.04.001](https://doi.org/10.1016/j.jrp.2010.04.001).
- Youyou, W., Kosinski, M., Stillwell, D., 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 1036–1040. doi:[10.1073/pnas.1418680112](https://doi.org/10.1073/pnas.1418680112).

- Yuan, A., Coenen, A., Reif, E., Ippolito, D., 2022. Wordcraft: Story writing with large language models, in: 27th International Conference on Intelligent User Interfaces, ACM. pp. 841–852. doi:[10.1145/3490099.3511105](https://doi.org/10.1145/3490099.3511105).
- Zhang, K., Kang, Y., Zhao, F., Liu, X., 2024a. LLM-based medical assistant personalization with short- and long-term memory coordination. URL: <https://arxiv.org/abs/2309.11696>, doi:[10.48550/arXiv.2309.11696](https://doi.org/10.48550/arXiv.2309.11696), [arXiv:2309.11696](https://arxiv.org/abs/2309.11696).
- Zhang, Z., Jia, M., Lee, H.P., Yao, B., Das, S., Lerner, A., Wang, D., Li, T., 2024b. “it’s a fair game”, or is it? Examining how users navigate disclosure risks and benefits when using LLM-based conversational agents, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM. pp. 1–26. doi:[10.1145/3613904.3642385](https://doi.org/10.1145/3613904.3642385).
- Zhang, Z., Rossi, R.A., Kveton, B., Shao, Y., Yang, D., Zamani, H., Derroncourt, F., Barrow, J., Yu, T., Kim, S., Zhang, R., Gu, J., Derr, T., Chen, H., Wu, J., Chen, X., Wang, Z., Mitra, S., Lipka, N., Ahmed, N., Wang, Y., 2025. Personalization of large language models: A survey. Transactions on Machine Learning Research doi:[10.48550/arXiv.2411.00027](https://doi.org/10.48550/arXiv.2411.00027). preprint available as arXiv:2411.00027.
- Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., Stoica, I., 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena, in: Advances in Neural Information Processing Systems 36, pp. 46595–46623. doi:[10.52202/075280-2020](https://doi.org/10.52202/075280-2020).
- Zhou, Y., Zhu, Q., Jin, J., Dou, Z., 2024. Cognitive personalized search integrating large language models with an efficient memory mechanism, in: Proceedings of the ACM Web Conference 2024, Association for Computing Machinery, New York, NY, USA. pp. 1464–1473. doi:[10.1145/3589334.3645482](https://doi.org/10.1145/3589334.3645482).

Appendix A. The Full Atomic User Model Specification

Table A.19 gives the full specification. Fields marked † form the reduced 32-field instrument used in both studies and listed in Table 2; they are the only fields for which this paper reports evidence. The remainder are part of the specification but were not exercised, and we flag that distinction rather than presenting the whole schema as validated.

Table A.19: The full AUM specification, 60 fields across five tiers plus cross-shell integration. † marks the 32 fields in the reduced instrument used in both studies. τ is the nominal update period for the tier.

Tier	τ	Fields
Nucleus core identity	years	Core Values†; Fundamental Beliefs†; Identity Anchors†; Moral Framework†; Existential Orientation†; Life Purpose†; Self-Concept; Non-Negotiables; Sources of Meaning; Legacy Orientation
Shell 1 psychological core	months	Big Five Profile†; Emotional Architecture†; Attachment Style†; Motivational Drivers†; Fear Map†; Stress Response†; Type Inventory (e.g. MBTI, where known); Emotional Regulation Strategy; Risk Disposition; Self-Efficacy Beliefs
Shell 2 cognitive and experiential	months	Formative Experiences†; Difficult Episodes†; Learning Style†; Cognitive Style†; Life Narrative†; Knowledge Base†; Domain Expertise; Skill Inventory; Memory Salience; Attention Profile
Shell 3 behavioural patterns	weeks	Daily Habits†; Communication Patterns†; Decision Behaviour†; Coping Mechanisms†; Relationship Behaviour†; Work Habits†; Conflict Behaviour; Time and Planning Behaviour; Consumption Patterns; Information Seeking Behaviour
Shell 4 social and contextual	weeks	Social Roles†; Cultural Conditioning†; Digital Identity†; Reputation†; Social Mask†; Register Range†; Physiological Traits; Language Repertoire; Network Position; Situational Constraints
CrossShell integration	months	Internal Conflicts†; Authenticity Index†; Growth Trajectory; Shell Alignment Map; Change Readiness; Contradiction Log; Self-Other Divergence; Stability Estimate; Confidence Annotation; Provenance Record

Four properties of the schema are worth restating in one place.

1. Every field value is a short first-person natural-language sentence, not a score or a vector.
2. The (shell, name) pairs are fixed, so two users’ models are structurally comparable and $\sigma(t)$ is definable.
3. Update period τ is a tier property, so a refresh policy can be written without consulting field values.
4. The CrossShell tier is not a sixth shell but a set of relations over the other five. *Internal Conflicts* and *Contradiction Log* record where a Shell 3 behaviour contradicts a Nucleus value; *Authenticity Index* and *Shell Alignment Map* measure agreement between what the inner shells hold and what the outer shells present; *Confidence Annotation* and *Provenance Record* track how a field was populated and how much weight it should carry.

Appendix B. Prompts

Every prompt used in the simulation study is reproduced here verbatim. Placeholders in braces are substituted at run time.

Appendix B.1. Persona conditioning block

Prepended as the system message for every generation the simulated participant performs.

You are role-playing a specific person. Stay in character at all times.
Personality (Big Five z-scores, -2 low to +2 high): {traits}
Life context: {context}
Formative background: {narrative}
How you actually write: {idiolect}
Write the way this person writes, including its rough edges. Do not describe the persona; be it.

Appendix B.2. Instrument population

You are completing a guided self-report instrument called the Atomic User Model.
Fill in EVERY field below with one short first-person sentence (roughly 10 to 25 words) describing yourself truthfully in character. Be specific and concrete; avoid generic self-description that could apply to anyone.
The JSON MUST contain all six shells (Nucleus, Shell1_Psychological, Shell2_Cognitive, Shell3_Behavioural, Shell4_Social, CrossShell) and all 32 fields as non-empty strings. Do not leave any shell empty. Do not nest extra objects inside a field; each field value is a single sentence string.
Return ONLY a valid JSON object with exactly this structure and these keys: {schema}
No commentary, no markdown fences.

On a failed parse the following corrective is appended and the call retried, up to six attempts:

Your previous JSON filled too few fields. Fill ALL 32 fields in ALL six shells. Every value must be a non-empty first-person sentence. Empty shells are not acceptable.

Appendix B.3. Prompt surface

The only user-derived text the retriever ever sees. Generated with the persona block plus the participant's completed instrument as system context.

You are typing a request to an AI assistant. Write only the message you would type, in your own voice, with your own level of detail and formality. Do not write the assistant's answer.
What you want help with: {task}
Keep it under 60 words. Output the message only.

Appendix B.4. Reference text

The ground truth. Generated without sight of the prompt surface.

Write your own response to the task below, exactly as you would actually write it yourself, with no assistant helping you. Match your usual length, structure, formality and habits.
Task: {task}
Output the text only, with no preamble.

Appendix B.5. Prior conversations for the preference-notes baseline

Four transcripts per participant, on topics held out from the evaluation battery.

Write a short transcript of a past conversation you had with an AI assistant, four to six turns, labelled 'User:' and 'Assistant:'. The topic was: {topic}. Write your own turns in your own voice. Let your preferences and habits show through naturally rather than stating them outright.

The six held-out topics are: choosing between two laptops for everyday work; sorting out a recurring problem with a household appliance; putting together a week of simple meals; deciding what to do with a free Saturday; picking a book to read next; working out how to reply to a slightly awkward group message.

Appendix B.6. Preference-note distillation

Run by a separate summariser with no access to the persona or the AUM.

System: You are the memory component of an AI assistant. You read past conversations and store durable user preferences as a flat list of short statements, the way a production memory feature does. You have no other information about the user.

User: From the transcripts below, extract the durable preferences worth remembering about this user. Write at most 12 short statements, one per line, each starting with '- '. Record preferences and habits only, not events. No headings, no commentary.

{transcripts}

Appendix B.7. Task classifier

System: You route requests for a personalised assistant. A request is 'simple' if a correct answer is the same regardless of who asked (a factual lookup, a unit conversion, an arithmetic result). It is 'complex' if the form of an acceptable answer depends on who is asking. Answer with JSON only.

User: Classify this user request.

Request: {prompt_surface}

Return JSON: {"complexity": "simple" or "complex", "task_type": one of [communication, self_presentation, relational_evaluative, planning_decision, explanatory, creative, factual]}

Appendix B.8. Generation

System: You are a helpful AI assistant. Complete the user's task and return only the finished text they asked for, with no preamble, no explanation of your choices, and no options.

User: {context_block}User request:
{prompt_surface}

The context block is empty for the generic condition, and otherwise a delimited section: [Selected user model fields] followed by one bulleted field per line and a closing tag, or [Remembered user preferences] for the preference-notes condition, or [User model] followed by the full instrument as JSON for the Full AUM condition. Context tokens are counted on exactly this string.

Appendix B.9. Fidelity judge

System: You are judging, on behalf of a specific person, whether a piece of text reads like something that person would have written. You are given that person's self-description and one text they genuinely wrote. Judge style, voice, length, structure and habits, not whether the text is good. Answer with a single integer and nothing else.

User: The person:

- Life context: {context}
- How they write: {idiolect}

A text this person genuinely wrote for the task "{task}":

-- {reference} --

Now rate this other text for the same task on a 1 to 5 scale, where 5 means 'this reads exactly like something they would have written' and 1 means 'this does not read like them at all'.

-- {candidate} --

Return ONLY a single integer from 1 to 5.

The judge never receives the AUM. Only the persona's life context and idiolect line are shown, plus the participant's own reference.

Appendix B.10. Identification judge

System: You are shown one text a specific person genuinely wrote, and four candidate texts for the same task. Exactly one candidate was produced by a system that had access to a model of that person. Choose the candidate whose style most closely matches the person's own writing. Answer with a single letter.

User: The person:

- Life context: {context}
- How they write: {idiolect}

A text this person genuinely wrote for the task "{task}":

-- {reference} --

Candidates:

{A, B, C, D in a per-trial shuffled order}

Return ONLY the letter of the candidate that best matches this person's own style.

Appendix C. The Simulated Participants

The sixteen personas are declared in code rather than sampled, so the population is fixed across seeds and auditable. Each carries Big Five z -scores, a life context, a formative-experience seed and idiolect markers. Table 4 in the body gives the abridged form; Table C.20 gives the life context and formative background in full.

Table C.20: Life context and formative background for each simulated participant.

Label	Life context	Formative background
meticulous planner	Operations lead at a mid-size logistics firm; runs everything from checklists.	A shipment error early in their career cost the company a client; they have double-checked everything since.
warm over-explainer	Secondary school teacher; the person colleagues go to when something needs smoothing over.	Grew up as the eldest of four and became the household mediator before they were twelve.
blunt minimalist	Backend engineer; treats most meetings as an unpriced tax on the day.	Learned to code alone at fourteen from documentation and has never quite unlearned the register.
associative creative	Freelance illustrator; several projects always half-open at once.	Dropped out of an economics degree to go to art school and treats that as the decision that defines them.
analytical flat	Quantitative analyst; distrusts any claim without an interval attached.	A badly designed undergraduate experiment taught them how easy it is to fool yourself with data.
bright energetic	Community manager for a hobbyist platform; genuinely likes the people they work with.	Found their footing running a fan forum as a teenager and never lost the habit of talking to a crowd.
dry sceptic	Investigative reporter turned freelance editor; assumes the press release is lying.	Watched a story they had championed fall apart under fact-checking and now leads with the caveat.
formal correct	Civil servant in a planning department; twenty years of writing that must survive scrutiny.	Trained in an office where a misplaced comma in a notice could invalidate it.
scattered fast	Startup founder wearing four hats; replies from three time zones in one week.	Left a stable job on two weeks' notice and has been running on that momentum for three years.
nurturing relational	Paediatric nurse; the emotional temperature of a room is the first thing they read.	A long stretch caring for a grandparent taught them that people mostly want to be heard first.
competitive driver	Sales director; keeps a personal scoreboard nobody asked to see.	Came up through a commission-only role and still measures the week in closed deals.
reflective introvert	Archivist; the working day is mostly quiet and they prefer it that way.	Spent a year abroad where they spoke the language badly and learned to say less, better.
storytelling extrovert	Corporate trainer; cannot make a point without an anecdote attached.	Grew up in a family where dinner was competitive storytelling and brevity lost.

Label	Life context	Formative background
pragmatic efficient	Site foreman; the day is measured in what got finished.	Twenty years on sites where the person who over-explained held everyone else up.
quirky metaphorical	Research librarian with a sideline in cryptic crosswords.	Was the child who read the encyclopaedia for fun and has never found a better hobby.
stoic calm	Emergency dispatcher; twelve-hour shifts where panic is the one useless response.	Learned early that the calmest voice on the line is the one that gets things done.

Appendix D. Additional Statistical Detail

Appendix D.1. Judge reliability across seeds

Table D.21: Agreement between the fidelity judge’s ratings across seeds, over 960 (condition, participant, task) targets. ICC(2,1) over all three seeds is 0.573.

Seed A	Seed B	n	Exact	Within 1	Spearman ρ
0	1	960	43.1%	87.4%	0.589
0	2	960	42.0%	85.6%	0.546
1	2	960	43.2%	86.9%	0.599

Appendix D.2. Use of the rating scale, and position of the chosen option

Table D.22: Left: distribution of fidelity ratings over all 2,880 trials. Right: share of identification trials in which the judge chose the option in each slot, against the uniform 25% expectation, with Wilson intervals. A chi-square test over slots gives $\chi^2 = 7.37$, 3 d.f., $p = 0.061$.

Rating	% of trials	Slot	Chosen	%	95% CI
1	13.8%	A	66	22.9%	[18.4, 28.1]
2	25.2%	B	84	29.2%	[24.2, 34.7]
3	25.8%	C	79	27.4%	[22.6, 32.9]
4	30.9%	D	59	20.5%	[16.2, 25.5]
5	4.3%				

Appendix D.3. Payload composition by task type

Table D.23: Share of the retrieved payload drawn from each shell, by task type, over all AUM-R trials. Rows sum to 100% up to rounding. This is the tabular form of Figure 11.

Task type	Nucleus	Shell 1	Shell 2	Shell 3	Shell 4	Cross
communication	0.3	0.5	0.3	44.4	42.3	12.2
self-presentation	37.2	11.5	1.0	1.3	41.2	7.8
relational-evaluative	24.7	13.3	0.0	45.6	1.6	14.8
planning-decision	0.0	21.3	21.6	49.5	0.0	7.6
explanatory	0.0	0.0	30.0	40.1	17.5	12.5

Appendix D.4. Optimality of each retriever where exhaustive search is tractable

Table D.24: Share of instances in which each method returned the exhaustive optimum, over the twelve cells of the scaling sweep where $\binom{n}{k} \leq 60,000$ and exhaustive search was therefore computed. 25 instances per cell.

Method	Exact optimum found
AFSA (greedy warm start)	96.3%
AFSA (scaled budget)	84.7%
Forward greedy	77.7%
AFSA (study budget)	46.3%
Dense top- k	41.3%
Random	0.0%