User-Centric Error Modeling Toward Cognitive Personalization in Language Model Systems

Published
February 6, 2026

Momen Ghazouani
Chief Scientist Setaleur Aplamda

“ True cognitive alignment is not measured by how well a model mimics our preferences, but by how accurately it navigates our rejections. We define our intellectual identity not just by what we seek, but by the specific boundaries of what we consider ***'error'****.”*

- *Momen Ghazouani Chief Scientist Setaleur Aplamda*

Abstract

This paper introduces **User-Centric Error Modeling (UCEM),** a conceptual framework that redefines personalization in language model systems. Rather than adapting to surface-level preferences or optimizing toward a singular notion of correctness, UCEM proposes that models should learn individualized definitions of error explicit, user-provided explanations of what constitutes an incorrect response relative to specific goals, reasoning patterns, domain assumptions, and working constraints. We argue that meaningful long-term personalization requires models to internalize user-specific error semantics through iterative feedback loops, moving beyond preference-based customization toward what we term **error-based cognitive personalization**. This paradigm positions users as active co-designers of their model's cognitive boundaries and raises fundamental questions about responsibility, epistemic alignment, and the nature of human-AI collaboration. We present UCEM not as a technical solution but as a theoretical repositioning of the personalization problem, outlining design principles, philosophical implications, scientific challenges, and open research questions necessary to operationalize this vision.

1. Introduction

The rapid advancement of large language models has transformed how humans interact with computational systems, enabling increasingly sophisticated forms of assistance across diverse domains. Yet despite progress in model capability, a persistent friction remains: models frequently produce outputs that are formally correct but contextually inappropriate, technically accurate but conceptually misaligned, or generically useful but specifically wrong for the individual user's needs. This friction is not primarily a matter of model competence but of personalization the capacity of a system to align its behavior with the idiosyncratic requirements, assumptions, and mental models of individual users.

Current approaches to personalization in language model systems predominantly operate at the level of preference: tone, verbosity, formatting style, content filters, or domain emphasis. These adaptations, while valuable, remain fundamentally superficial. They adjust **how** information is presented without fundamentally altering **what the model considers correct** in the context of a specific user's cognitive framework. A model may learn that a user prefers concise responses in markdown format, but it does not learn that this user defines **"correct reasoning"** differently in probabilistic contexts, operates under domain-specific assumptions that contradict textbook knowledge, or considers certain inference patterns erroneous within their particular application constraints. This paper proposes a conceptual shift from preference-based personalization to what we term **error-based cognitive personalization**. The core proposition is deceptively simple yet foundational: language models should learn not only what users like, but what users consider **wrong** and more importantly, *why* they consider it wrong. These explanations of error, provided explicitly by users in response to model outputs, constitute a form of personalized ground truth that reflects individual cognitive boundaries, reasoning norms, domain-specific correctness criteria, and task-specific constraints.

We introduce **User-Centric Error Modeling (UCEM)** as a theoretical framework for instantiating this shift. In UCEM, users do not merely rate outputs or select preferences from predefined categories. Instead, they engage in an ongoing process of error specification: explaining to the model why particular responses fail relative to their own goals and mental models. These explanations accumulate into personalized error representations that guide future model behavior, creating a closed-loop learning relationship in which the model progressively converges toward cognitive compatibility with each user rather than toward a universal standard of correctness.

This paradigm fundamentally reframes the user's role. Rather than passive consumers of model outputs who can only accept, reject, or regenerate, users become partial co-designers of their model's cognitive boundaries. This elevation of user agency carries profound implications: it necessitates new forms of responsibility for shaping and maintaining personalized models, raises questions about epistemic ownership and the distribution of cognitive labor, and challenges conventional assumptions about what it means to align AI systems with human values and needs.

The contribution of this work is explicitly conceptual rather than technical. We do not present algorithms, architectures, or implementation strategies as finalized solutions. Instead, we aim to redefine the problem space of personalization itself, arguing that current framings are insufficiently ambitious and that the research community should direct attention toward deeper forms of cognitive alignment. The paper explores what such alignment might entail, what challenges it would pose, and what questions would need to be answered to make it tractable.

The remainder of this paper proceeds as follows. Section 2 situates UCEM within existing personalization paradigms and articulates the conceptual distinctions that motivate this work. Section 3 presents the core framework of User-Centric Error Modeling, introducing key concepts including user-defined error semantics, personalized error representations, and the closed-loop learning process. Section 4 examines the philosophical and ethical dimensions of this paradigm, particularly concerning responsibility, agency, and epistemic

authority. Section 5 outlines scientific and engineering challenges that would need to be addressed, framed as open research questions rather than solved problems. Section 6 critically examines limitations, risks, and potential failure modes of the UCEM paradigm. Section 7 discusses future research directions, and Section 8 concludes .

2. Personalization Paradigms: From Preferences to Errors

2.1 The Current Landscape of Model Personalization

Personalization in language model systems currently manifests across several dimensions, each operating at different levels of abstraction and making different assumptions about what constitutes individual alignment. At the most basic level, session-based adaptation uses conversation history to maintain coherence and reference earlier exchanges, but this is ephemeral and resets with each new interaction. More persistent approaches include user preference settings that specify desired output characteristics length, formality, technical depth, language that are applied uniformly across interactions.

Recent developments have introduced more sophisticated personalization mechanisms. Some systems maintain user profiles that encode stated preferences, domain interests, or role-specific requirements. Others employ feedback signals explicit ratings, regeneration requests, or binary approval indicators to fine-tune responses or adjust recommendation algorithms. Retrieval-augmented approaches allow models to access user-specific documents or knowledge bases, enabling contextually grounded responses that incorporate individual information.

These methods represent meaningful progress, yet they share a common limitation: they personalize the presentation and retrieval of information without fundamentally altering the model's underlying conception of correctness. A model may learn that a particular user prefers technical explanations with mathematical notation, but it does not learn that this user employs a Bayesian framework that treats probability fundamentally differently from frequentist approaches, or that within their research domain, certain widely-accepted assumptions are locally rejected. The model adapts its style and information selection but not its epistemic framework.

This limitation becomes particularly acute in domains where correctness is not universal but context-dependent, assumption-laden, or contested. In scientific research, different theoretical frameworks produce incompatible but internally consistent explanations. In professional practice, industry-specific norms and regulatory constraints define what counts as appropriate reasoning. In creative domains, individual aesthetic principles and project-specific requirements determine what constitutes a valid contribution. Current personalization mechanisms lack the representational capacity to capture these deeper forms of individual variation in what constitutes error.

2.2 The Insufficiency of Prompt Engineering

A natural objection to the need for deeper personalization mechanisms is that existing tools particularly sophisticated prompt engineering already enable users to specify their requirements with arbitrary precision. If a user can articulate their assumptions, constraints, and reasoning preferences in a detailed prompt, what further personalization is necessary?

This objection misunderstands both the nature of the problem and the limits of explicit specification. First, many aspects of individual cognitive frameworks are tacit rather than explicit. Users often cannot fully articulate their reasoning patterns, assumption structures, or error criteria in advance; these emerge through interaction, becoming visible only when the model produces outputs that violate them. The process of recognizing and explaining errors is itself a form of knowledge externalization that cannot be reduced to upfront specification.

Second, the cognitive burden of comprehensive prompt engineering is prohibitive for sustained interaction. Requiring users to fully specify their epistemic framework in every prompt transforms each interaction into an exercise in formal specification. This is not only inefficient but fundamentally misaligned with natural human communication, which relies heavily on shared context, implicit assumptions, and progressive refinement through feedback.

Third, prompt engineering operates at the level of individual queries rather than persistent adaptation. Each prompt is interpreted independently, with no accumulation of learned error patterns across interactions. A user might repeatedly encounter the same category of error perhaps the model consistently applies inappropriate probability distributions in statistical reasoning and must repeatedly engineer prompts to prevent it, because the model has no mechanism for learning that this category of reasoning is wrong for this particular user.

The distinction is between **specification And learning**. Prompt engineering asks users to specify requirements; UCEM proposes that models should learn user-specific definitions of error through accumulated feedback. This shift is analogous to the difference between writing explicit rules for every possible situation and providing examples from which general patterns are induced. The former requires complete foresight and constant vigilance; the latter enables progressive improvement through interaction.

2.3 From Preferences to Errors: A Conceptual Distinction

The central conceptual move of UCEM is the shift from modeling user preferences to modeling user-defined errors. This distinction requires clarification, as both preferences and errors constitute forms of evaluative feedback that indicate alignment or misalignment between model output and user requirements.

Preferences express what users like or want: stylistic choices, presentation formats, content emphases, or output characteristics. They are fundamentally additive indicating desired features to include and typically continuous, admitting degrees of satisfaction. A user may prefer more technical language or less formal tone, with gradations of acceptability.

Preferences are often context-dependent but not necessarily principled; they may shift based on mood, task, or circumstance without constituting inconsistency. Errors, in contrast, express what users consider **incorrect** relative to their goals, constraints, and mental models. They are fundamentally subtractive indicating violations to avoid and often binary, marking clear boundaries between acceptable and unacceptable outputs. More importantly, errors typically reflect principled commitments: domain-specific knowledge, methodological assumptions, reasoning patterns, or constraint structures that define correctness within a particular cognitive framework.

This distinction has profound implications for learning and generalization. Preference learning seeks to match user desires and can tolerate approximation; error learning seeks to avoid violations and requires precision about boundaries. A model that learns a user prefers concise explanations can generate responses of varying lengths with decreasing satisfaction. A model that learns a user considers frequentist statistical reasoning erroneous in Bayesian contexts must recognize and avoid this entire category of response, as any instance constitutes failure.

The shift to error-based personalization also implies a different relationship between user and model. Preference-based systems position users as consumers expressing tastes to be satisfied; error-based systems position users as teachers providing corrective feedback about cognitive boundaries. The former is transactional; the latter is developmental. This developmental dimension becomes central to UCEM's vision of long-term cognitive alignment.

2.4 Cognitive Personalization as the Target State

If preference-based personalization operates at the surface level of output characteristics, cognitive personalization operates at the deeper level of reasoning patterns, assumption structures, and domain-specific correctness criteria. The goal is not merely to adjust how information is presented but to align the model's internal processes with the user's cognitive framework what we term **cognitive compatibility.**

Cognitive compatibility does not mean that the model replicates human reasoning or achieves mutual intelligibility in its internal representations. Rather, it means that the model's outputs consistently conform to user-specific definitions of correct reasoning, valid inference, appropriate assumption structures, and acceptable constraint satisfaction. The model may achieve this through mechanisms entirely unlike human cognition, but the functional outcome is alignment at the epistemic rather than merely stylistic level.

This target state raises immediate questions about feasibility, desirability, and risk. Can language models learn such individualized cognitive frameworks? Should they? What happens when user-defined correctness criteria conflict with broader social norms, ethical principles, or factual accuracy? These questions are not resolved here but are central to the research agenda that UCEM proposes. Before addressing them, we must first articulate what the paradigm entails in more concrete terms.

3. The UCEM Framework : Core Concepts and Mechanisms

3.1 User-Defined Error Semantics

“ Preference captures the style of the answer; error modeling captures the architecture of the reasoning. To move from a tool to a partner, the system must learn not merely to speak in our voice, but to think within our constraints.”

- *Momen Ghazouani Chief Scientist Setaleur Aplamda*

The foundational concept of UCEM is that errors are not objective failures of model capability but subjective violations of user-specific correctness criteria. What constitutes an error for one user may be perfectly acceptable for another, even given identical inputs and contexts. This subjectivity is not arbitrary but structured by individual differences in goals, domain knowledge, methodological commitments, working constraints, and reasoning styles.

User-defined error semantics refers to the explicit explanations users provide about why a model output is incorrect relative to their particular framework. These explanations go beyond simple rejection signals they articulate **what** is wrong and **why** it is wrong. For example, rather than marking a response as unhelpful, a user might explain: "This probability calculation uses a frequentist approach, but in my research framework, probability represents degree of belief, not long-run frequency, so this reasoning is fundamentally incorrect for my purposes."

These explanations serve multiple functions. Pedagogically, they teach the model about user-specific correctness criteria. Representationally, they provide structured information that can be encoded into personalized error models. Epistemically, they externalize tacit knowledge that users hold about their own domains and methods but may not have previously articulated. The act of explaining errors forces users to make explicit their often-implicit cognitive boundaries.

The scope of user-defined error semantics is deliberately broad. It encompasses domain-specific knowledge errors (violations of specialized facts or theories), methodological errors (use of inappropriate analytical frameworks), assumption errors (reliance on premises the user rejects), constraint errors (violation of task-specific requirements), and reasoning errors (application of inference patterns the user considers invalid in particular contexts). This breadth reflects the diversity of ways in which model outputs can fail to align with individual cognitive frameworks.

Critically, user-defined error semantics need not align with global notions of correctness, consensus knowledge, or model training objectives. A researcher might define certain widely-accepted statistical practices as erroneous within their heterodox theoretical framework. A professional might reject textbook solutions due to industry-specific regulatory constraints. A creative practitioner might consider conventional approaches fundamentally wrong relative to their artistic vision. UCEM takes these definitions seriously as valid personalization targets, while acknowledging that doing so raises complex questions about epistemic authority and the limits of individualized alignment.

3.2 Personalized Error Representations

User-defined error semantics must be encoded in forms that enable models to generalize from specific corrections to broader patterns. This requires constructing what we term **personalized error representations** structured knowledge about what categories of outputs this particular user considers erroneous and under what circumstances.

The nature of these representations is an open research question, but conceptually they must capture several dimensions of information. First, they must encode the **content** of errors: what specific facts, methods, assumptions, or reasoning patterns are considered incorrect. Second, they must encode **context dependencies** : when and under what conditions these errors apply, as correctness criteria may shift across domains, tasks, or interaction modes. Third, they must encode **explanatory rationales** : why these outputs are considered erroneous, linking errors to underlying principles or constraints that govern user-specific correctness.

One can envision multiple possible instantiations of personalized error representations, each with different properties. They might take the form of explicit rules or constraints that outputs must satisfy or avoid. They might be embedded representations that cluster error explanations in semantic space, enabling similarity-based retrieval. They might be structured knowledge graphs that link error categories to contexts and rationales. They might be fine-tuned model components that have learned to avoid certain output patterns. The technical realization is less important than the functional requirement: the representation must enable the model to recognize and avoid errors as defined by each specific user.

A crucial property of personalized error representations is their cumulativeness. Unlike session-based context that resets after each interaction, error representations should persist and accumulate across the user's history with the system. Each new error explanation refines, extends, or qualifies the existing representation. Over time, this accumulation ideally converges toward a comprehensive model of the user's cognitive boundaries what they consider correct and incorrect, under what circumstances, and according to what principles.

This cumulative property introduces questions about representation capacity, maintenance, and evolution. How much personalized error information can be effectively stored and utilized? How should contradictory error definitions be reconciled? How can representations adapt as user needs and frameworks evolve? These challenges are discussed further in Section 5.

3.3 The Closed-Loop Learning Process

UCEM instantiates a fundamentally different interaction dynamic than conventional language model systems. Rather than one-shot query-response cycles, UCEM envisions a closed-loop learning process in which user and model engage in iterative refinement of cognitive alignment. The process begins when the model produces an output that the user identifies as erroneous. Rather than simply regenerating or moving on, the user provides an explicit error explanation articulating what is wrong and why. This explanation is incorporated into the user's personalized error representation. In subsequent interactions, the model attempts to avoid similar errors by consulting this representation when generating

responses. When errors still occur as they inevitably will, particularly early in the relationship the user provides additional explanations that further refine the representation. Over many interactions, the model progressively learns the boundaries of user-specific correctness, reducing the frequency of errors and increasing cognitive compatibility.

This closed-loop structure has several important characteristics. First, it is **bidirectional** : the model's outputs elicit user corrections, which modify the model's future behavior, which in turn affects what outputs are generated and what further corrections are needed. Second, it is **progressive :** each cycle ideally increases alignment, reducing error rates and deepening the model's understanding of user-specific correctness criteria. Third, it is **ongoing :** there is no predetermined endpoint at which learning ceases; the process continues as long as the user-model relationship persists.

The closed-loop framing emphasizes the relational nature of personalization in UCEM. Cognitive compatibility is not a static property achieved through initial configuration but an emergent outcome of sustained interaction. The quality of alignment depends on both the model's learning capacity and the user's engagement in providing error explanations. This shared responsibility for maintaining the relationship distinguishes UCEM from systems that place the burden of adaptation entirely on the model.

The learning process also implies a temporal dimension often absent from current personalization approaches. Early interactions are characterized by higher error rates and more frequent user corrections as the model learns the user's framework. Over time, as the error representation becomes more comprehensive, the model should require less corrective feedback, entering a maintenance phase where errors are rare and corrections primarily address edge cases or evolving requirements. This temporal arc transforms the user experience from one of constant correction to one of reliable cognitive partnership.

3.4 Users as Co-Designers of Cognitive Boundaries

The shift to error-based personalization fundamentally redefines the user's role in the system. In preference-based paradigms, users are consumers who select from predefined options or provide feedback that guides optimization. In UCEM, users become active co-designers of their model's cognitive boundaries the principles and constraints that define correct behavior within their personalized framework.

“ In the UCEM paradigm, the user ceases to be a mere consumer of intelligence and becomes the curator of its validity. We are not just prompting for answers; we are authoring the epistemic boundaries of our own digital extensions.”

- *Momen Ghazouani Chief Scientist Setaleur Aplamda*

This co-design role manifests in several ways. Most directly, users design by specifying: through error explanations, they explicitly define what the model should and should not do, effectively writing the rules that govern personalized behavior. They design by curation: through selective correction, they determine which errors warrant explanation and which can be tolerated, shaping the error representation's priorities. They design by refinement: through iterative feedback, they progressively clarify and extend their cognitive framework,

discovering and articulating boundaries they may not have recognized in advance. The co-designer framing carries implications for agency, authority, and expertise. It positions users as experts on their own requirements an assumption that may be warranted in professional, creative, or research domains but potentially problematic in contexts where users lack domain knowledge or hold misconceived frameworks. It grants users epistemic authority over personalized correctness criteria, even when these diverge from consensus knowledge or model training objectives. It assumes users possess the metacognitive capacity to identify and explain errors in their own terms, which may not hold across all populations or contexts.

This elevation of user agency is both empowering and demanding. It empowers users to shape AI systems according to their genuine needs rather than conforming to predetermined notions of correct usage. It respects individual cognitive diversity and legitimizes alternative frameworks that might be marginalized in one-size-fits-all approaches. But it also demands significant cognitive labor identifying errors, formulating explanations, maintaining consistency that not all users may be willing or able to provide. The co-designer role is not passive consumption but active participation in system design.

3.5 The User Memory Space Concept

To operationalize personalized error representations, UCEM requires persistent storage dedicated to individual users. We introduce the concept of a **user memory space** : a dedicated allocation of system resources for maintaining user-specific information, analogous to user-allocated storage in consumer computing devices.

The user memory space would house personalized error representations along with any other information necessary for sustained cognitive alignment: accumulated error explanations, context dependencies, evolving correctness criteria, interaction history relevant to learning, and metadata about the user's domains, goals, and frameworks. This space is owned by and attributable to the specific user, persisting across sessions and interactions, and potentially portable across different instances or deployments of the model system.

The memory space concept raises important design questions. What should be the capacity constraints? How should information be organized for efficient retrieval? What mechanisms should govern what enters the memory space automatic logging of all error explanations, user-controlled curation, intelligent filtering? How should conflicts between old and new information be resolved? What privacy and security protections should apply?

The memory space also introduces questions about ownership and control. If the space contains user-provided error explanations that reflect their professional expertise, creative vision, or research frameworks, does the user own this intellectual property? Can users export, delete, or transfer their memory spaces? What happens to the space if the user stops using the system? These questions intersect with broader debates about data ownership and user rights in AI systems.

Perhaps most fundamentally, the memory space concept makes concrete the resource costs of deep personalization. Just as personal computing devices allocate storage to individual users, UCEM requires allocating computational and storage resources to individual

personalization. This allocation is not infinite; practical implementations would face constraints on memory capacity, retrieval efficiency, and computational overhead. These constraints would in turn shape what kinds of personalization are feasible and at what scale. We return to these practical challenges in Section 5.

4. Philosophical and Ethical Dimensions

4.1 Responsibility and the Distribution of Cognitive Labor

The UCEM paradigm fundamentally redistributes responsibility between user and system in ways that deserve careful examination. In conventional language model interactions, responsibility for output quality rests primarily with the model developer: through training, safety measures, and capability improvements, developers strive to ensure models produce helpful, harmless, and honest outputs. Users may provide feedback, but their primary responsibility is simply to use the system appropriately.

UCEM disrupts this distribution by making users partially responsible for shaping their model's behavior. If a personalized model produces errors, the question arises: whose error is it? If the model violates a user-specific correctness criterion that was never explained to it, the failure may lie with the user's incomplete specification rather than the model's inadequacy. If the model correctly follows a user-defined error pattern that happens to be misguided, responsibility for the resulting harm is shared between the user who specified the pattern and the system that enabled this specification.

This shared responsibility is not necessarily problematic indeed, it may be a more honest reflection of how cognitive tools function in practice. Calculators do not bear sole responsibility for correct arithmetic; users must input appropriate numbers and operations. Databases do not bear sole responsibility for query results; users must formulate valid queries. Similarly, deeply personalized language models might not bear sole responsibility for outputs; users must provide adequate error guidance. The question is whether this distribution of responsibility is transparent, equitable, and sustainable.

The cognitive labor required for UCEM participation is non-trivial. Users must identify errors in model outputs, formulate coherent explanations, maintain consistency across corrections, and monitor the model's learning progress. This labor is skilled work requiring metacognitive awareness, domain expertise, and communicative ability. It is reasonable to ask whether this labor should be required for effective AI use, or whether it creates unfair burdens that advantage already-privileged users who possess the time, expertise, and confidence to engage in sophisticated feedback. There is also a temporal dimension to responsibility. Early in a user-model relationship, when the error representation is sparse, the model will make frequent mistakes that require correction. This period demands sustained user engagement and tolerance for imperfect performance. Users must essentially train their personalized models through patience and repeated feedback. This investment may yield long-term benefits in cognitive compatibility, but it front-loads costs in ways that may deter adoption or exclude users unable to make such investments.

4.2 Epistemic Authority and the Limits of Personalization

UCEM grants users epistemic authority to define correctness within their personalized frameworks, even when these definitions diverge from consensus knowledge, scientific consensus, or factual accuracy. This raises profound questions about the limits of personalization: how far should systems defer to user-defined correctness criteria?

Consider a user who rejects evolution, climate science, or vaccine safety based on ideological commitments. Should a personalized model learn to treat evolutionary explanations, climate warnings, or vaccine recommendations as errors? Or should there be boundaries beyond which user-defined correctness cannot override factual accuracy or scientific consensus? If such boundaries exist, who determines them, and on what basis? One response is that UCEM should be domain-limited: personalization might apply to subjective domains (aesthetic judgments, methodological preferences, project-specific constraints) but not to objective facts or well-established scientific knowledge. This boundary is appealing but difficult to operationalize. Many domains exist in grey areas where consensus is contested, evidence is incomplete, or multiple valid frameworks coexist. Statistics offers competing philosophical foundations; economics encompasses incompatible theoretical schools; history involves interpretive frameworks that yield different accounts of the same events. Should personalization apply in these contested spaces?

Another response is that epistemic authority should be distributed: users can define correctness for their purposes, but models should retain the ability to flag divergences from consensus knowledge or to present alternative frameworks. A model might learn that a user rejects frequentist statistics but still note when a particular problem is conventionally approached from a frequentist perspective. This preserves user agency while maintaining epistemic transparency.

A more radical position is that epistemic authority properly belongs to users in personalized contexts. If a professional operates under industry-specific regulatory frameworks that contradict textbook approaches, the textbook approaches are indeed errors within that professional context. If a researcher employs a heterodox theoretical framework, outputs conforming to orthodox frameworks are genuinely incorrect for their research purposes. Correctness is context-dependent, and personalization means respecting the contexts that matter to users, even when these depart from general norms. These positions are not easily reconciled. They reflect deeper tensions in epistemology about the nature of knowledge, the authority of expertise, and the relationship between individual and collective understanding. UCEM does not resolve these tensions but makes them concrete and actionable, forcing system designers to make explicit choices about epistemic boundaries.

4.3 Identity, Evolution, and Cognitive Commitment

Personalization based on accumulated error history creates a form of cognitive commitment: the model learns to behave according to the user's framework, which becomes encoded in the error representation. This commitment raises questions about identity and evolution. As users grow, learn, and change, their cognitive frameworks evolve. What happens when a user's current error definitions conflict with their past specifications? A user might begin their relationship with a model as a graduate student employing one theoretical framework,

then shift to a different framework as a researcher, then adopt yet another perspective later in their career. Each transition potentially invalidates error definitions from previous periods. Should the model automatically adapt to these shifts, or should it maintain consistency with established error patterns? Should users be able to explicitly reset their error representations, or should accumulated learning persist?

These questions touch on fundamental issues of personal identity and continuity. If a personalized model reflects a user's cognitive framework at a particular time, does it continue to represent that user when their framework fundamentally changes? Is there a stable core of user-specific correctness that persists across cognitive evolution, or is personalization necessarily time-bound and revisable?

One approach is to version error representations, maintaining historical snapshots that users can revert to or reference. Another is to implement gradual adaptation mechanisms that detect shifts in error patterns and slowly update representations. A third is to make all personalization explicitly temporary and revisable, with users retaining full control over what persists.

The identity question also relates to authenticity. Does deep personalization risk creating echo chambers in which models only reflect back what users already believe, never challenging assumptions or introducing contrary perspectives? If a model learns that certain reasoning patterns are errors for a user, does it stop presenting those patterns entirely, even when they might be valuable for growth, learning, or critical self-reflection? The relationship between personalization and intellectual development is complex and potentially contradictory.

4.4 Privacy, Transparency, and User Control

User memory spaces containing personalized error representations would constitute rich sources of information about individual users their expertise, beliefs, assumptions, working constraints, cognitive patterns, and domain knowledge. This information is potentially sensitive, revealing not just preferences but the deeper structure of how users think and reason.

The privacy implications are significant. Error explanations might expose proprietary methodologies, confidential project constraints, or personal epistemic commitments that users would not want disclosed. If these representations are stored by service providers, they could be subject to data breaches, legal requests, or commercial exploitation. Even if protected from external access, they might be used internally to profile users, target advertising, or train general models in ways that violate user expectations. UCEM therefore requires strong privacy protections and user control mechanisms. Users should have transparency about what information is stored in their memory spaces, how it is used, and who has access. They should retain the right to inspect, modify, export, or delete their error representations. They should be able to control whether their personalized data is used for any purpose beyond direct personalization of their own experience.

Transparency about model behavior is equally important. Users should understand how their error explanations influence model outputs, what limitations exist in the

personalization system, and when the model is uncertain about how to apply their correctness criteria. Opaque personalization where users cannot understand why the model behaves as it does or how their feedback has been incorporated undermines the co-design relationship that UCEM envisions.

These transparency and control requirements are challenging to implement while maintaining the benefits of personalization. Fully interpretable error representations might sacrifice expressiveness or learning efficiency. Complete user control over representations might enable harmful manipulations or inconsistent specifications. There are tensions between privacy protection and the computational requirements of personalization that require careful navigation.

5. Scientific and Engineering Challenges

This section outlines major challenges that would need to be addressed to operationalize UCEM. These are presented not as solved problems but as open research questions that define the agenda for developing this paradigm.

5.1 Learning from Heterogeneous Error Feedback

How can models effectively learn from the diverse and often sparse error explanations that users provide? User feedback will vary enormously in quality, specificity, consistency, and technical sophistication. One user might provide detailed, systematic error analyses with clear rationales; another might offer terse, ambiguous corrections. One user might maintain consistent correctness criteria across thousands of interactions; another might be internally contradictory or change positions without explanation.

Learning mechanisms must be robust to this heterogeneity. They must extract meaningful patterns from sparse and noisy signals, distinguish between genuine correctness criteria and momentary preferences, handle contradictory specifications gracefully, and adapt to different feedback modalities (natural language explanations, structured annotations, implicit signals). The challenge is compounded by the fact that each user's error feedback is statistically sparse: there may be limited examples of any particular error category, making generalization difficult.

Moreover, error feedback is fundamentally different from the supervision signals used in model training. It is negative rather than positive, specifying what to avoid rather than what to produce. It is contextual and conditional rather than universal. It may reference abstract principles rather than concrete examples. Standard supervised learning paradigms may be inadequate for incorporating such feedback effectively.

5.2 Representation and Retrieval of Personalized Error Knowledge

What form should personalized error representations take, and how should they be efficiently retrieved and applied during generation? The representation must balance multiple competing requirements: expressiveness sufficient to capture diverse error categories, compactness for efficient storage and retrieval, interpretability for user understanding and control, and computational efficiency for real-time application during

model inference. Candidate representations might include explicit rule sets, embedding spaces, structured knowledge graphs, or model fine-tuning. Each has distinct tradeoffs. Explicit rules are interpretable but may lack coverage and flexibility. Embeddings are compact but opaque. Knowledge graphs are structured but complex to maintain. Fine-tuning is powerful but computationally expensive and difficult to control. Hybrid approaches that combine multiple representation types might be necessary but introduce additional complexity.

Retrieval presents its own challenges. During generation, the model must identify which error patterns are relevant to the current context, retrieve the corresponding constraints, and apply them without excessive computational overhead. This requires efficient indexing, context-sensitive retrieval mechanisms, and integration with the generation process that does not degrade response quality or latency. The retrieval problem becomes more difficult as error representations grow over time, potentially encompassing thousands of user-specific patterns.

5.3 Generalization Across Contexts and Abstraction Levels

How should models generalize from specific error corrections to broader patterns? When a user explains that a particular statistical approach is incorrect in one context, should the model avoid that approach entirely, only in similar contexts, or only when explicitly matching specific conditions? Overgeneralization risks applying corrections too broadly, creating new errors; undergeneralization fails to leverage error feedback efficiently, requiring repeated corrections.

The generalization problem is complicated by the multiple levels at which errors can be defined. Some errors are concrete: "Do not use this specific formula in this specific situation." Others are categorical: "Do not use frequentist reasoning in Bayesian contexts." Still others are principled: "Do not make assumptions that contradict my theoretical framework." The model must learn not only the specific corrections but the appropriate level of abstraction and scope of application.

Context-sensitivity adds another layer of complexity. A reasoning pattern that is an error in one domain or task might be perfectly appropriate in another. The model must learn when correctness criteria apply and when they do not, potentially requiring sophisticated context representations and conditional error models. The boundary between legitimate context-dependence and inconsistent specification is not always clear.

5.4 Handling Inconsistency and Contradiction

Users are not perfectly consistent. They may provide contradictory error definitions, change their minds without acknowledgment, or apply different standards in different contexts without realizing it. How should systems handle such inconsistencies? Should they attempt to reconcile contradictions, flag them for user resolution, prioritize recent over old specifications, or maintain multiple potentially inconsistent error models?

Contradiction detection itself is nontrivial, requiring the system to recognize when two error definitions are genuinely incompatible rather than context-specifically applicable. Once detected, there are several possible responses. The system might query the user for clarification, but this could be intrusive and burdensome. It might attempt to infer resolution based on temporal ordering or confidence levels, but this risks making incorrect assumptions. It might simply maintain contradictory specifications and apply context-specific heuristics, but this could lead to unpredictable behavior.

There is also the question of when inconsistency reflects error versus legitimate evolution. A user's changing error definitions might indicate growth, learning, or shifting contexts rather than confusion. Distinguishing between these cases requires understanding user intent and development, which is itself a challenging inference problem.

5.5 Computational and Scalability Constraints

Deep personalization at the level UCEM envisions would impose significant computational costs. Each user requires dedicated storage for error representations, personalized retrieval during generation, and potentially user-specific model adaptations. As user bases grow and error representations accumulate, these costs multiply. What are the practical limits on personalization depth given realistic computational budgets?

There are fundamental tradeoffs between personalization depth and scalability. Richer error representations enable better cognitive alignment but require more storage and computational resources. More sophisticated learning mechanisms improve from feedback more effectively but increase processing overhead. Longer interaction histories provide more context for personalization but strain memory and retrieval systems.

System designers will need to make difficult choices about how to allocate limited resources across users and over time. Should all users receive equal personalization capacity, or should resource allocation scale with usage? Should older error patterns be gradually forgotten to free resources, or should they persist indefinitely? Should personalization be tiered, with deeper customization available only to users willing to pay computational costs? These decisions have significant implications for equity and accessibility.

5.6 Evaluation and Validation

How should personalized error modeling be evaluated? Standard model evaluation metrics perplexity, accuracy against benchmark datasets, human preference ratings are inadequate for assessing cognitive compatibility with individual users. Success in UCEM is inherently subjective and user-specific: the model should produce outputs that conform to each user's correctness criteria, which may vary arbitrarily across users. Evaluation might focus on error reduction rates: does the frequency of user-identified errors decrease over time? But this metric is complicated by changing user standards, evolving tasks, and the possibility that users become less vigilant about correction as they habituate to the system. Alternatively, evaluation might assess consistency: does the model reliably apply learned error patterns? But consistency without accuracy could simply mean reliably implementing incorrect constraints.

User satisfaction is an obvious candidate metric, but it confounds multiple factors: personalization quality, general model capability, interface usability, and user expectations. Moreover, satisfaction may not align perfectly with cognitive compatibility; users might be satisfied with outputs that do not fully conform to their stated error criteria, or dissatisfied with correct personalization for unrelated reasons.

Validation faces the fundamental challenge that ground truth for user-specific correctness does not exist outside the user's own judgments. There are no external benchmarks against which to measure personalized alignment. This raises concerns about verification, auditability, and quality assurance that are difficult to address within the UCEM paradigm.

6. Limitations, Risks, and Failure Modes

6.1 Overfitting to Idiosyncratic or Harmful Frameworks

Perhaps the most serious risk of UCEM is that models might overfit to user frameworks that are idiosyncratic, misguided, or actively harmful. A model that learns to treat well-established scientific facts as errors because a user rejects them could reinforce dangerous misinformation. A model that adapts to racist, sexist, or otherwise bigoted correctness criteria becomes an instrument of harm. A model that internalizes a user's conspiracy theories or pseudoscientific beliefs could amplify their negative societal effects.

The personalization-safety tradeoff is fundamental to UCEM. The more completely a model adapts to user-defined correctness, the more it risks abandoning broader safety, accuracy, or ethical constraints. Yet restricting personalization to preserve these constraints contradicts the core premise of UCEM that users should be able to define correctness for their own purposes.

Potential mitigations include maintaining baseline constraints that cannot be overridden by personalization, implementing transparency mechanisms that flag divergences from consensus knowledge, or requiring justification for error definitions that contradict established facts. But each mitigation weakens the paradigm's central commitment to user epistemic authority. There may be no perfect resolution to this tension, only different ways of balancing competing values.

6.2 Cognitive Lock-in and Reduced Exploration

Deep personalization might create cognitive lock-in, where models become so adapted to user frameworks that they cease to present alternative perspectives, challenge assumptions, or introduce novel ideas. If a model learns that certain reasoning patterns or information sources are errors, it may stop surfacing them entirely, even when they might be valuable for user growth, learning, or creative exploration.

This risk is particularly acute for users whose frameworks are self-reinforcing or whose error definitions systematically exclude disconfirming information. The model becomes an echo chamber, reflecting back only what aligns with existing beliefs and filtering out everything else. Over time, this could narrow rather than expand user capabilities, constraining rather than augmenting cognition.

The tension here is between respecting user agency their right to define correctness for themselves and promoting intellectual development and exposure to diverse perspectives. A model that constantly violates user correctness criteria is frustrating and unhelpful; one that never challenges them may be comforting but intellectually stifling. Finding the right balance requires sophisticated understanding of when to conform to user error definitions and when to thoughtfully deviate.

6.3 Exclusion and Inequality of Access

The cognitive labor required for effective UCEM participation may create new forms of exclusion. Users with less education, technical sophistication, time, or confidence may struggle to provide the detailed error explanations necessary for deep personalization. This could result in a two-tiered system where advantaged users receive highly personalized, cognitively compatible models while others receive generic, poorly aligned experiences.

Moreover, the benefits of personalization accumulate over time through sustained interaction. Users who can afford to invest in long-term relationships with their models reap increasing returns, while those who interact sporadically or cannot commit to ongoing feedback receive minimal benefit. This temporal dimension of advantage compounds existing inequalities in access to cognitive tools.

There is also the question of domain expertise. UCEM works best when users possess the metacognitive awareness and domain knowledge to identify and explain errors accurately. In domains where users are novices or where error identification requires expertise they lack, personalization may be ineffective or even counterproductive, potentially reinforcing misconceptions rather than supporting learning.

6.4 Privacy Breaches and Misuse

User memory spaces containing personalized error representations would be valuable targets for various forms of misuse. Competitors might seek access to professionals' error explanations to reverse-engineer proprietary methodologies. Adversaries might exploit error patterns to craft targeted manipulations or social engineering attacks. Governments or institutions might compel disclosure to infer users' beliefs, political positions, or professional activities.

Even well-intentioned use of personalized data raises concerns. Aggregating error patterns across users could enable profiling, discrimination, or differential treatment based on inferred characteristics. Using personalized data to train general models could leak individual information or create models that preferentially serve some users over others.

The sensitivity of error explanations is often not obvious until one considers what they reveal. A series of corrections about appropriate statistical methods exposes a user's training and professional framework. Error explanations about topic sensitivity reveal personal boundaries and potentially traumatic experiences. Corrections about factual matters expose what a user does and does not know, creating asymmetric information that could be exploited.

6.5 Maintenance Burden and Relationship Decay

UCEM envisions ongoing user-model relationships that require sustained maintenance. Users must periodically review their error representations, update them as their needs evolve, reconcile contradictions, and provide feedback on whether personalization remains aligned with their goals. This maintenance burden may be sustainable initially but could become onerous over time, particularly as error representations grow complex. Relationship decay is a plausible failure mode: users begin with enthusiasm, invest heavily in training their models, but gradually disengage as the novelty wears off or as the maintenance burden accumulates. Without ongoing feedback, personalization stagnates or becomes misaligned as user needs change. The model retains error patterns that are no longer relevant while failing to learn new ones, resulting in a system that is neither generic enough to be broadly useful nor personalized enough to provide real benefit.

This decay might be particularly pronounced during transitions: new jobs, career changes, shifting research interests, or life events that alter user contexts and requirements. The personalization accumulated under previous circumstances becomes obsolete, but users may lack the time or energy to rebuild error representations for new contexts. The result is a system stuck in an outdated configuration that no longer serves the user well.

7. Open Research Directions

7.1 Formal Models of User-Specific Correctness

A foundational challenge is developing formal frameworks for representing user-specific correctness criteria. What mathematical or logical structures can capture the diverse ways users define errors? How can these structures support reasoning about consistency, generalization, and context-dependence? Can we develop type systems, constraint languages, or logical frameworks that enable users to specify correctness criteria at appropriate levels of abstraction?

Research might draw on diverse sources: formal verification and constraint satisfaction for representing correctness conditions, epistemic logic for modeling user knowledge and beliefs, preference learning for understanding user values and goals, and cognitive science for understanding how humans conceptualize and explain errors. Integrating insights from these domains could yield principled foundations for user-centric error modeling.

7.2 Interactive Explanation and Feedback Mechanisms

How should systems elicit error explanations from users? What interaction modalities make error specification natural, efficient, and accurate? Research could explore various approaches: natural language explanations, contrastive examples showing correct versus incorrect outputs, interactive labeling of specific error components, or collaborative debugging where user and model jointly identify sources of misalignment.

There are also questions about feedback granularity and selectivity. Should users explain every error or only significant patterns? Should systems request explanations proactively when detecting potential errors, or wait for user initiative? How can feedback mechanisms minimize cognitive burden while maximizing information value? Designing interactions that make error specification feel natural rather than burdensome is crucial for UCEM adoption.

7.3 Temporal Dynamics of Personalization

The temporal evolution of user-model relationships deserves systematic study. How do error patterns accumulate and change over time? What characterizes successful versus unsuccessful personalization trajectories? Are there critical periods when user investment is most valuable? How should systems balance continuity (maintaining learned patterns) with adaptability (updating to changing requirements)? Longitudinal research could track users through different phases of relationship development: initial training, deepening alignment, maintenance, evolution, and potentially disengagement. Understanding these phases could inform design of support mechanisms tailored to each stage, from onboarding assistance in early phases to adaptive maintenance in later periods.

7.4 Hybrid Approaches Combining Multiple Personalization Levels

UCEM need not replace existing personalization mechanisms but could complement them in hybrid systems that operate at multiple levels simultaneously. Surface-level preference personalization could coexist with deep error-based cognitive personalization, with the former handling stylistic adaptation and the latter handling epistemic alignment. Research could explore how to effectively integrate multiple personalization mechanisms: when to apply which level, how to prevent conflicts between different personalization types, how to allocate resources across mechanisms, and how to present unified interfaces that do not expose unnecessary complexity. The goal would be personalization systems that are both accessible for casual use and powerful for deep cognitive alignment.

7.5 Collaborative and Organizational Personalization

This paper has focused on individual user personalization, but many cognitive activities are collaborative or organizational. Research teams share methodological frameworks, organizations maintain institutional knowledge, and professional communities develop domain-specific standards. Can UCEM extend to collective error modeling where groups jointly define and maintain shared correctness criteria?

Collaborative personalization introduces new challenges: reconciling different individuals' error definitions within shared contexts, managing permissions and authority over collective error representations, handling evolution as group membership changes, and enabling both individual and collective levels of personalization to coexist. But it also offers new possibilities for creating shared cognitive tools that embody collective expertise and standards.

7.6 Relationship Between UCEM and Model Capabilities

How does user-centric error modeling interact with underlying model capabilities? Does deeper personalization require more capable base models, or can sophisticated error modeling compensate for limited capabilities? Are some model architectures more amenable to personalization than others? How do personalization and capability improvement trade off against each other in system design?

There may be complementarities between capabilities and personalization: better models might learn from error feedback more effectively, while personalization might make limited models more useful by aligning them precisely to user needs. Or there might be tensions: resources devoted to personalization infrastructure might be better spent improving base capabilities. Understanding these dynamics would inform resource allocation and development priorities.

8. Conclusion

This paper has introduced User-Centric Error Modeling as a conceptual framework for rethinking personalization in language model systems. The central proposition is that meaningful personalization requires moving beyond preference-based customization to error-based cognitive alignment, where models learn individualized definitions of correctness through explicit user feedback about what constitutes errors relative to specific goals, constraints, and frameworks.

UCEM represents a shift in how we conceive the relationship between users and AI systems. Rather than passive consumers of model outputs, users become active co-designers of their model's cognitive boundaries. Rather than optimizing toward universal notions of correctness, models progressively converge toward cognitive compatibility with individual users. Rather than one-shot interactions, personalization emerges through sustained closed-loop learning relationships.

This paradigm raises profound questions that have been explored throughout this paper. Philosophically, it challenges conventional distributions of responsibility and epistemic authority, positioning users as partial owners of model behavior while raising questions about the limits of personalization and the risks of overfitting to harmful frameworks. Scientifically, it poses difficult challenges around learning from heterogeneous feedback, representing and retrieving personalized knowledge, generalizing across contexts, and evaluating inherently subjective alignment. Practically, it demands significant computational resources, sophisticated interaction mechanisms, and ongoing user engagement that may create new forms of inequality.

We have not attempted to resolve these questions or to present UCEM as a finalized solution. Instead, we have offered it as a theoretical repositioning of the personalization problem one that we believe the research community should seriously consider as language models become more deeply integrated into individual cognitive work. The current paradigm of preference-based personalization, while valuable, is insufficient for the kinds of cognitive partnership that many users will require in professional, creative, research, and complex decision-making contexts.

Whether UCEM is the right framework for addressing this insufficiency remains an open question. The risks and limitations we have discussed are real and substantial. There may be fundamental barriers to operationalizing this vision at scale, or insurmountable tensions between personalization and safety that make the paradigm untenable. Alternative approaches that achieve similar goals through different mechanisms may prove more tractable.

But we believe the questions this framework raises are worth pursuing, even if the specific answers differ from what we have proposed. How can AI systems move beyond surface-level adaptation to deeper forms of cognitive alignment? What role should users play in defining correctness for their own purposes? How can we build systems that respect individual cognitive diversity while maintaining appropriate bounds on what can be personalized? What mechanisms enable sustained learning relationships between humans and AI that grow more aligned over time?

These questions will only become more pressing as language models increase in capability and integration with human work. The gap between what these systems can do and how well they align with individual users' actual needs will likely widen without more sophisticated personalization paradigms. UCEM offers one possible direction for closing this gap not through better training, larger models, or more data, but through fundamentally rethinking what it means for an AI system to be aligned with an individual user. The research agenda opened by this work is extensive. It requires contributions from multiple disciplines: machine learning and natural language processing for the technical mechanisms, human-computer interaction for the interface design, cognitive science for understanding human error conception, philosophy for the epistemic and ethical dimensions, and social science for studying the societal implications. It requires both theoretical development and empirical investigation, conceptual clarity and practical experimentation. Most fundamentally, it requires a willingness to question the assumption that personalization is primarily about preferences and to consider the possibility that it might need to be about something deeper the structure of thought, the boundaries of correctness, and the frameworks through which individuals make sense of information and make decisions. Whether this deeper form of personalization is desirable, feasible, or wise remains to be determined. But asking the question seriously seems essential for a future in which AI systems serve not just as general-purpose tools but as genuine cognitive partners.

Acknowledgments

This is a conceptual and theoretical position paper proposing a research direction rather than reporting completed work. The ideas presented are intended to stimulate discussion and investigation rather than to claim definitive solutions. The author acknowledges that many of the questions raised remain unresolved and that the framework proposed may require substantial revision as understanding develops.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Fogarty, J., Akumu, J., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The TAMER framework. Proceedings of the Fifth International Conference on Knowledge Capture, 9-16.

Ouyang, L., Han, J., Jiang, X., Carrillo, G., Wang, J., Fehlwagner, T., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.